



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Doctoral Thesis

Pathway and Network Analysis of
Transcriptomic and Genomic Data

Sora Yoon

Department of Biological Sciences

Graduate School of UNIST

2019

Pathway and Network Analysis of Transcriptomic and Genomic Data

Sora Yoon

Department of Biological Sciences

Graduate School of UNIST


Pathway and Network Analysis of Transcriptomic and Genomic Data

A thesis/dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Sora Yoon

12/31/2018 of submission

Approved by



Advisor

Dougou Nam

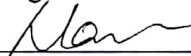
Pathway and Network Analysis of Transcriptomic and Genomic Data

Sora Yoon

This certifies that the thesis/dissertation of Sora Yoon is approved.

12/11/2018 of submission

signature



Advisor: Prof. Dougu Nam

signature



Prof. Yun Joo Yoo

signature



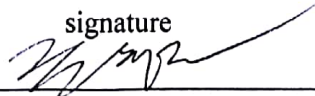
Prof. Cheol-Min Ghim

signature



Prof. Semin Lee

signature



Prof. Taejoon Kwon

Abstract

The development of high-throughput technologies has enabled to produce omics data and it has facilitated the systemic analysis of biomolecules in cells. In addition, thanks to the vast amount of knowledge in molecular biology accumulated for decades, numerous biological pathways have been categorized as gene-sets. Using these omics data and pre-defined gene-sets, the pathway analysis identifies genes that are collectively altered on a gene-set level under a phenotype. It helps the biological interpretation of the phenotype, and find phenotype-related genes that are not detected by single gene-based approach. Besides, the high-throughput technologies have contributed to construct various biological networks such as the protein-protein interactions (PPIs), metabolic/cell signaling networks, gene-regulatory networks and gene co-expression networks. Using these networks, we can visualize the relationships among gene-set members and find the hub genes, or infer new biological regulatory modules.

Overall, this thesis/dissertation describes three approaches to enhance the performance of pathway and/or network analysis of transcriptomic and genomic data. First, a simple but effective method that improves the gene-permuting gene-set enrichment analysis (GSEA) of RNA-sequencing data will be addressed, which is especially useful for small replicate data. By taking absolute statistic, it greatly reduced the false positive rate caused by inter-gene correlation within gene-sets, and improved the overall discriminatory ability in gene-permuting GSEA. Next, a powerful competitive gene-set analysis tool for GWAS summary data, named GSA-SNP2, will be introduced. The z-score method applied with adjusted gene score greatly improved sensitivity compared to existing competitive gene-set analysis methods while exhibiting decent false positive control. The performance was validated using both simulation and real data. In addition, GSA-SNP2 visualizes protein interaction networks within and across the significant pathways so that the user can prioritize the core subnetworks for further mechanistic study. Finally, a novel approach to predict condition-specific miRNA target network by biclustering a large collection of mRNA fold-change data for sequence-specific targets will be introduced. The bicluster targets exhibited on average 17.0% (median 19.4%) improved gain in certainty (sensitivity + specificity). The net gain was further increased up to 32.0% (median 33.2%) by filtering them using functional network information. The analysis of cancer-related biclusters revealed that PI3K/Akt signaling pathway is strongly enriched in targets of a few miRNAs in breast cancer and diffuse large B-cell lymphoma. Among them, five independent prognostic miRNAs were identified, and repressions of bicluster targets and pathway activity by mir-29 were experimentally validated. The BiMIR database provides a useful resource to search for miRNA regulation modules for 459 human miRNAs.

Table of Contents

Abstract	i
Table of Contents.....	iii
List of Figures.....	vi
List of Tables	viii
Chapter I: Introduction.....	1
1.1 Omics data	1
1.1.1 Genomic data	1
1.1.1.1 Genome-wide association study.....	2
1.1.2 Transcriptomic data.....	4
1.1.2.1 Microarray and RNA-sequencing.....	4
1.1.2.2 Issues in RNA-sequencing data analysis	5
1.2 Pathway analysis	6
1.2.1 Pathway databases.....	7
1.2.2 Pathway analysis methods	7
1.2.2.1 Over-representation analysis.....	7
1.2.2.2 Functional class sorting	7
1.2.2.2.1 Gene-set enrichment analysis.....	8
1.2.2.3 Pathway topology-based method.....	9
1.2.3 Competitive and self-contained gene-set analysis.....	9
1.3 Biological network	12
1.4 Research overview	13
Chapter II: Improving gene-set enrichment analysis of RNA-seq data with small replicates	15
2.1 Abstract.....	15
2.2 Introduction.....	15
2.3 Materials and Methods	16
2.3.1 Absolute gene-permuting GSEA and filtering.....	16
2.3.2 Simulation of the read count data with the inter-gene correlation.....	18
2.3.3 Biological relevance measure of a gene-set	20
2.3.4 RNA-seq data handling and gene-set condition	21
2.3.5 Gene-set size	21

2.3.6 AbsfilterGSEA R package	21
2.4 Results	21
2.4.1 Comparison of gene-permuting GSEA methods for simulated read count data	21
2.4.2 Comparison of GSEA methods for RNA-seq data.....	25
2.4.3 Effects of the absolute filtering on false positive control and biological relevance	26
2.5 Discussion.....	27
2.6 Supplementary information of Chapter II.....	29

Chapter III: A powerful pathway enrichment and network analysis tool for GWAS summary data 34

3.1 Abstract.....	34
3.2 Introduction.....	34
3.3 Materials and Methods	36
3.3.1 Algorithm of GSA-SNP2.....	36
3.3.2 Competitive pathway analysis tools	38
3.3.3 Simulation study.....	38
3.4 Results and Discussion	39
3.4.1 Type I error rate simulation test	39
3.4.2 Power simulation test.....	39
3.4.3 Performance comparison using real data	41
3.4.4 Comparison of competitive and self-contained pathway analysis results.....	46
3.4.5 Comparison with the competitive pathway analysis for gene expression data	46
3.4.6 Network visualization.....	48
3.5 Conclusion.....	49
3.6 Supplementary information of Chapter III	50

Chapter IV: Biclustering analysis of transcriptome big data identifies condition-specific miRNA targets 67

4.1 Abstract.....	67
4.2 Introduction.....	67
4.3 Materials and Methods	70
4.3.1 Collection of expression fold-change data.....	70
4.3.2 Sequence-specific miRNA targets.....	70
4.3.3 miRNA target prediction using a Progressive Bicluster Extension (PBE) algorithm.....	70
4.4 Results	72
4.4.1 Comparison with existing biclustering algorithms.....	72
4.4.2 Accuracy of the biclustering target prediction	73

4.4.3	Comparison with anticorrelation-based methods in cancer	75
4.4.4	miRNAs targeting PI3K/Akt signaling in cancer	78
4.4.5	BiMIR: a bicluster database for condition-specific miRNA targets.....	80
4.5	Discussion.....	81
4.6	Supplementary information of Chapter IV	83
Chapter V: Discussion and conclusion.....		112
References		113
Acknowledgement (감사의 글).....		133

List of Figures

Figure 1.1. SNP association test using Chi-squared test and effect size evaluation.....	3
Figure 1.2. Comparison of cDNA microarray and RNA-sequencing.....	5
Figure 2.1. The relationship between the mixing coefficient (α) and the average inter-gene correlation.....	20
Figure 2.2. Performance comparison of gene-permuting GSEA methods for simulated read counts. .	22
Figure 2.3. Average receiver operating characteristic (ROC) curves.	24
Figure 3.1. The monotone cubic spline trend curves.....	36
Figure 3.2. Type 1 error rate comparison.....	40
Figure 3.3. Statistical power comparison.....	41
Figure 3.4. Power comparison using real data.	44
Figure 3.5. Comparison of gene p-value distributions in the pathways that are only significant with (a) GSA-SNP2 or (b) sARTP.....	46
Figure 3.6. PPI network (HIPPIE) from DIAGRAM data.....	47
Figure 4.1. Two approaches for miRNA regulation module discovery.....	69
Figure 4.2. Overview of the biclustering-based miRNA target prediction.	71
Figure 4.3. Simulation test for biclustering algorithm.	73
Figure 4.4. Performance of miRNA target prediction using binding sequence, biclustering, and functional networks.....	75
Figure 4.5. Performance comparison between biclustering and anticorrelation-based methods.....	77
Figure 4.6. miRNA targets in PI3K/Akt pathway (breast cancer).....	79

Supplementary Figures

Figure S2.1. Performance comparison of gene-permuting GSEA methods for simulated read counts.	29
Figure S2.2. Average receiver operating characteristic (ROC) curves for two sample cases.	30
Figure S2.3. The effect of absolute gene-permuting GSEA.....	32
Figure S3.1. Dual cubic spline illustration.....	54
Figure S3.2. Power comparison using real data with strict significance cutoff.	66
Figure S4.1. Progressive bicluster extension (PBE) algorithm.	86
Figure S4.2. Pseudocode of Progressive Bicluster Extension.....	87
Figure S4.3. Distribution of the number of conditions, genes and density in biclusters with three different fold-change cut-offs.	90
Figure S4.4. ESC/iPSC biclusters searched by multiple biclustering methods.....	94
Figure S4.5. microRNA targets in PI3K/Akt pathway (DLBCL).	105
Figure S4.6. BiMIR database.	111

List of Tables

Table 1.1. Types of genetic variation	2
Table 1.2. Size factor of six between-sample normalization methods.....	6
Table 1.3. Popular pathway databases	8
Table 1.4. Competitive and self-contained gene-set analysis methods for gene expression data.	10
Table 1.5. Competitive and self-contained gene-set analysis methods for GWAS summary data.....	11
Table 2.1. Significant gene-sets detected by the absolute GSEA-GP filtering (FDR<0.1) with the mod- t score (DHT-treated and control LNCaP cell line).	26
Table 3.1. Power comparison using canonical pathways for diabetes.....	45
Table 3.2. Running times for seven pathway analysis programs for GWAS summary data.	48

Supplementary Tables

Table S3.1. Gene Ontology terms (mSigDB C5 v6.0) related to 15 height-related categories.....	58
Table S3.2. Canonical pathways (mSigDB C2 v6.0) related to 15 T2D-related pathways	64
Table S4.1. Existing miRNA target prediction tools	83
Table S4.2. Statistics of BiMIR biclusters.....	91
Table S4.3. Real data analysis.....	95
Table S4.4. Let-7c bicluster targets regulating pluripotency or up-regulated in ES/iPS cells.	98
Table S4.5. The accuracy for bicluster targets of eleven test miRNAs	100
Table S4.6. The accuracy of 1.3-fold bicluster targets filtered by node degree.....	101
Table S4.7. The accuracy of 1.5-fold bicluster targets filtered by node degree.....	102
Table S4.8. The accuracy of 2.0-fold bicluster targets filtered by node degree.....	103
Table S4.9. microRNA expression patterns in cancers reported from the literature	104
Table S4.10. Functional enrichment test for miR-29, miR-34a, miR-145 targets in DLBCL	106
Table S4.11. Functional enrichment test for miR-1, miR-29, miR-34a, miR-145 targets in breast cancer	107
Table S4.12. Multivariate Cox regression analysis of microRNAs in the DLBCL dataset	108
Table S4.13. Multivariate Cox regression analysis of microRNAs in the breast cancer dataset.....	109

Chapter I: Introduction

1.1 Omics data

A suffix ‘-ome’ represents the mass of something, and it is frequently used to indicate a group of biological molecules. For example, genome, transcriptome and proteome represent the complete sets of DNA, transcripts (RNA) and proteins in a cell, respectively. With the development of high-throughput technology, it has become possible to produce such omics data within short time. It facilitates the systematic analysis of genetic and/or epigenetic features of diseases and helps to find therapeutic and diagnostic targets. Here, the concepts and characteristics of genomic (especially for GWAS data) and transcriptomic data will be explained.

1.1.1 Genomic data

In a broad sense, the genomic data refers to any data come from genome of such as nucleotide sequences, annotations or read alignments. Among them, I will focus on the genetic variation data in this thesis/dissertation. Many diseases are caused by genetic variations (Table 1.1). The variants within coding region may alter the protein structure, and those in the non-coding regulatory region can affect to the gene expression regulation. The genomic variants are classified into two groups based on the variant size. One is the simple nucleotide variation (SNV) including single nucleotide polymorphism (SNP) and short insert/deletion. Another is the structural variation (SV) including long insert/deletion, copy number variation (CNV), inversion and translocation. Table 1.1 describes the definition and example diseases of each variant type.

The genome-wide profiling of human genetic variations has been possible with the construction of human reference genome ¹ and two great projects such as International HapMap Projects ² and 1000 Genome Project ³ that produced reference haplotype data for human genetic variations. In the International HapMap Project, more than 3 million human common SNPs had been genotyped for 1,301 individuals from 11 populations (Phase III), and identified about 500,000 tag SNPs that represent the behaviors of each linkage disequilibrium (LD) block. In 1000 Genome Project, whole genome sequencing (WGS) had been done for 2,504 individuals from 26 population, and discovered an extensive number of genomic variants including 84.7 million SNPs, 3.6 million indels and 60,000 structural variants ⁴. These reference haplotype panels are great sources for genome-wide association study that facilitate an efficient genotyping through imputation, which will be explained in the next section in detail.

Table 1.1. Types of genetic variation

Variation Type	Description	Example disease
Simple Nucleotide Variation (SNV)		
Single nucleotide polymorphism (SNP)	Single nucleotide variation found more than 1% of population. More than 84 millions of SNPs have been found in human genome.	Sickle-cell anaemia ⁵ Wilson's disease ⁶ Tay-Sachs disease ⁷
Indel	Insertion and deletion of base pairs (length: 1~10,000 bp). 1.6~2.5 millions of indels are found in human genome.	Cystic fibrosis ⁸ Fragile X syndrome ⁸ muscular dystrophy ⁸
Satellite	Repetition of DNA motifs, typically 5-50 times. <ul style="list-style-type: none"> • microsatellites (< 10 bp per repeat) • minisatellites (10–60 bp per repeat) • satellites (~hundreds bp per repeat) • macrosatellites (several kb per repeat) 	Huntington's disease ⁹ Fragile X syndrome ⁹ Myotonic dystrophy ⁹
Structural variation (SV)		
Copy number variation (CNV)	Copy number change of long DNA segment (>1 kb)	Huntington's disease ¹⁰ Alzheimer disease ¹¹ Autism ¹²
Inversion	Rearrangement of DNA segment to reverse orientation.	Haemophilia A ¹³
Translocation	Rearrangement of DNA segment to be inserted into different chromosome	Leukaemia ¹⁴ Ewing's sarcoma ¹⁵

1.1.1.1 Genome-wide association study

The genome-wide association study (GWAS) is carried out to identify the genetic variants (mainly SNPs) that are associated with a phenotype (e.g., disease)¹⁶. For example, if one type of allele of a SNP is more frequently observed in patient group compared to the control group, the SNP is regarded as a marker of the disease of interest. The higher SNP effect size (represented by odds ratio for dichotomous trait or beta value for continuous trait; figure 1.1) represents the stronger association of that SNP with the phenotype. The phenotype can be either dichotomous trait where the samples consist of case and control groups (e.g., disease vs. normal) or continuous trait such as height and BMI. For dichotomous phenotype, Chi-squared test for independence is widely performed to evaluate a SNP's association p-value. Figure 1.1 represents the process of association test using chi-squared test for a SNP and calculating its effect size. Logistic regression is an alternative method to perform the association test. It is used to adjust various confounders such as ethnicity and batch effect. For quantitative traits, following linear regression model is used to perform association test.

$$Y = \beta_0 + \beta_1 X + \sum_{i=0}^N \gamma_i C_i$$

where, Y is a phenotype vector, X is a (normalized) genotype vector of a SNP, C_i is confounding factors, β_0 is intercept, β_1 is regression coefficient of X and γ_i is regression coefficient of confounding factor C_i . Here, β_1 represents the effect size of the SNP.

Although there are more than 84 million SNPs in human genome, we don't need to perform association test for all of them. As mentioned in the previous section, the International HapMap Project, launched in 2003, identified ~500,000 human tag SNPs that represent each haplotype. In typical GWAS, these tag SNPs are genotyped first using SNP array to find significantly associated tag SNPs (e.g., $p < 5 \times 10^{-8}$). Next, association test is performed again for all SNPs in the haplotypes of significant tag SNPs. In this step, the unknown genotypes are inferred from reference panel constituted from International HapMap Project or 1000 Genome project. This step is called imputation. It enables to find more accurate SNP marker without genotyping all SNPs. SNP markers found in this discovery stage are often further validated using independent cohort.

The SNP markers identified in various phenotypes can be referred from GWAS Catalog page (<https://www.ebi.ac.uk/gwas/>). As of December 2018, 89,251 unique SNP-trait associations ($p\text{-value} < 5 \times 10^{-8}$) are reported in GWAS Catalog.

The result of association test for all SNPs is often provided with summarized format including columns of SNP ID, genomic position, effect allele, effect size and association p-value. This kind of data is called 'GWAS summary data'. It is favorably used for further pathway analysis due to its relatively small data size.

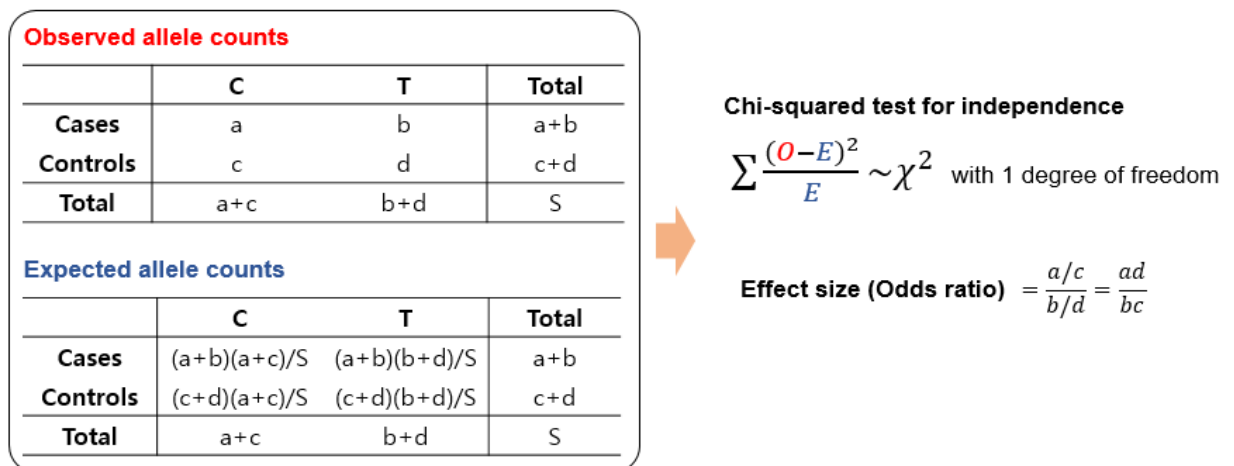


Figure 1.1. SNP association test using Chi-squared test and effect size evaluation.

The tables represent the observed (denoted as O) and expected (denoted as E) SNP variant counts in case and control samples. The association p-value of the SNP is evaluated using Chi-squared test for independence. The effect size of the SNP is calculated by odds ratio of variant counts between case and control samples.

1.1.2 Transcriptomic data

Various protein coding- and non-coding RNAs are transcribed from DNA in a cell. The transcriptome means the entire RNA molecules in a cell, but it usually indicates the entire set of specific RNA type of interest such as messenger RNAs (mRNAs). In this thesis/dissertation, the transcriptome represents a complete set of mRNAs. Among all RNAs, the majority is composed of ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) (95~97%), while the mRNAs that we mainly focus on occupy merely less than 5%¹⁷. Thus, the enrichment of mRNAs (or other RNA type of interest) or depletion of rRNA and tRNA is carried out after RNA extraction. The purified mRNA expression levels are measured by cDNA microarray or RNA-sequencing, and this transcriptomic data is used to measure (1) the expression level of transcripts in a specific condition, (2) alternative splicing to predict the isoform protein levels and (3) the effect of genomic variants on gene expression¹⁸⁻²¹.

1.1.2.1 Microarray and RNA-sequencing

Similar to genomic data, the transcriptomic data is measured using (1) cDNA microarray or (2) RNA-sequencing (RNA-seq). The differences of two methods are represented in figure 1.2. The cDNA microarray was developed a decade earlier than RNA-seq (The first studies of cDNA microarray and RNA-seq were published in 1995 and 2008, respectively²²⁻²³). It measures all known transcripts' levels at the same time based on the hybridization. Although useful, there are two limitations in this method. First, it can't discover novel transcripts because the probes on a microarray chip are produced only for known transcripts. Second, it shows high background noises caused by nonspecific hybridization between transcripts and probes²⁴. The RNA-seq technique was developed to solve these problems. By aligning the RNA fragments to the reference genome, it can discover de-novo RNA molecules²⁵. In addition, it shows quite low background noise and high sensitivity to detect lowly expressed RNA molecules. In spite of these advantages of RNA-seq, there are several things to be careful in analyzing RNA-seq data as described in below.

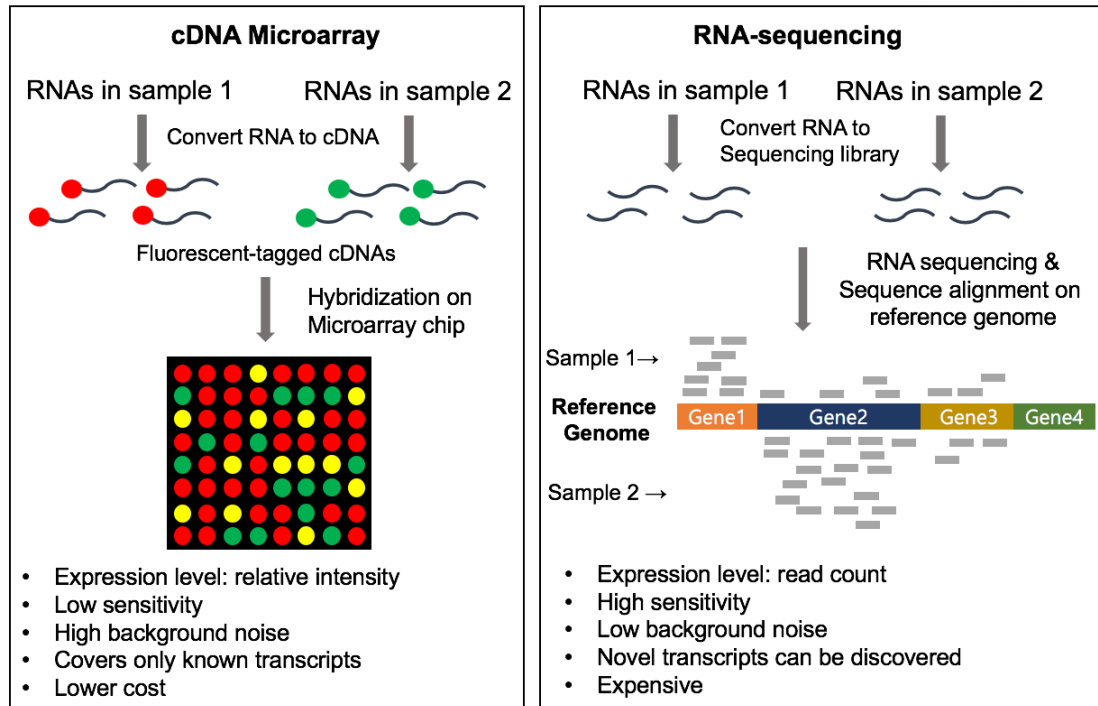


Figure 1.2. Comparison of cDNA microarray and RNA-sequencing

1.1.2.2 Issues in RNA-sequencing data analysis

The issues in analyzing RNA-seq data arise from its expression measuring method (counting the number of reads aligned on each gene). First issue is the normalization. It is a process to make the expression levels comparable within a sample or between samples. For example, the raw read counts of gene A and gene B within a sample cannot be directly compared because longer genes tend to be mapped with more reads. To remove this gene length bias, the raw counts of gene A and B are typically normalized by RPKM ($= \frac{\text{Raw read count} \times 10^9}{\text{gene length} \times \text{library size}}$) or FPKM ($= \frac{\text{Fragment count} \times 10^9}{\text{gene length} \times \text{library size}}$). The gene length bias is not considered when comparing the gene expression levels between samples (e.g., differential expression analysis of a gene). Instead, the ‘sequencing depth bias’ must be corrected in this case. Many ‘between-sample’ normalization methods such as DESeq²⁶, TMM²⁷ or UQ had been devised considering the library size factor. Table 1.2 represents how each method normalizes the raw read counts. Another issue is the statistical evaluation of differential expression. Because RNA-seq read counts are discrete values, Poisson distribution had been used in the early days. However, the assumption of Poisson distribution (mean and variance are same) was not fit to the real RNA-seq data where the gene count variances are often much larger than gene count means. Thus, over-dispersed Poisson distribution, a.k.a. Negative Binomial (NB) distribution, have been frequently used in modeling RNA-seq data. In NB distribution, the variance of a i -th gene in the j -th sample (σ_{ij}^2) is defined as the sum of expected mean (μ_{ij}) and an additional term.

$$\sigma_{ij}^2 = \mu_{ij} + \varphi_i \mu_{ij}^2$$

Here, φ_i is the dispersion coefficient of i -th gene. The size of dispersion coefficient depends on the data type. For example, the dispersion coefficient of a dataset consisting of samples from unrelated individuals (e.g., cancer cohort data) will be much higher than that of those consisting of technical replicates or genetically identical samples (e.g., cell lines). Many RNA-seq DE analysis methods such as DESeq2²⁸, edgeR²⁷, baySeq²⁹ and EBSeq³⁰ use the negative binomial model. Voom³¹ is another DE analysis method that transforms the normalized read counts to log-scale and applies the linear model which is commonly used in the microarray analysis. There are also non-parametric methods such as NOISeq³² or SAMseq³³.

Table 1.2. Size factor of six between-sample normalization methods.

For each method, raw RNA-seq read count of gene g in j -th sample (K_{gj}) is normalized by dividing it with size factor of j -th sample (s_j). G and m represent the total number of genes and samples, respectively. $UQ(x)$ is upper-quartile value of x , and Q_j is the upper-quartile count in j -th sample.

Methods	Size factor of j -th sample
Total count method (TC)	$s_j = \frac{\sum_{g=1}^G K_{gj}}{\sum_{g=1}^G \sum_{j=1}^m K_{gj}}$
Upper-Quartile method (UQ)	$s_j = UQ\left(\frac{K_{gj}}{\sum_{g=1}^G K_{gj}}\right)$
Median method (Med)	$s_j = median\left(\frac{K_{gj}}{\sum_{g=1}^G K_{gj}}\right)$
Quantile-normalization (Q)	$s_j = 10^{\log_{10} Q_j - (\frac{1}{m}) \sum_{v=1}^m \log_{10} Q_v}$
TMM	$\log_2(s_j) = \frac{\sum_{g \in G'} w_{gj} M_{gj}}{\sum_{g \in G'} w_{gj}}$
Where $M_{gj} = \log_2((K_{gj}/N_j)/(K_{gr}/N_r))$, $w_{gj} = \frac{N_j - K_{gj}}{N_j K_{gj}} + \frac{N_r - K_{gr}}{N_r K_{gr}}$, N_j , N_r are the total number of reads for j -th sample and reference sample r , respectively. G' is set of genes not trimmed by fold change and average expression level cutoff.	
DESeq	$median\left(\frac{K_{gj}}{(\sum_{v=1}^m K_{gv})^{1/m}}\right)$

1.2 Pathway analysis

One basic approach to analyze these omics data is to identify the list of genes significantly altered between case and control groups. Such analysis is called differential expression (DE) analysis for

transcriptomic, and GWAS for genomic data. This gene-based analysis has been widely performed to find various disease-causing genes, and extended our biological knowledges.

In addition, the pathway-based (gene-set-based) analysis provides useful information. An organism maintains its life through the complex interactions among numerous biological pathways. Dysregulation in some metabolic pathways can lead to chronic diseases or even cancers³⁴. The pathway analysis is performed to find the genetic difference between case and control groups on gene-set level. There are several advantages in the pathway analysis. First, it helps easier interpretation of the common biological function of the significantly altered genes (e.g., DE genes), especially when numerous genes are significantly detected. Second, it improves the reproducibility of signature genes among independent studies³⁵. Third, it reduces the multiple correction burden and increases the detection power, especially for GWAS data³⁶.

1.2.1 Pathway databases

To perform pathway analysis, a list of pre-defined pathways is required. The Pathguide database (<http://pathguide.org>) provides links to 702 pathway databases and their information³⁷. Among them, 251 were those of human pathway databases. Those databases are classified into 10 categories (protein-protein interactions, metabolic pathways, signaling pathways, transcription factors/ gene regulatory networks, protein-compound interactions, genetic interaction networks, protein sequence focused and others). Table 1.3 describes 13 popular pathway databases.

1.2.2 Pathway analysis methods

The pathway analysis is classified into three types based on the gene-set scoring method as follows.

1.2.2.1 Over-representation analysis

From the omics data with two sample groups, we typically identify differentially altered genes between groups using a significance cutoff (e.g., $FDR < 0.05$). Let say such genes are signature genes. The over-representation analysis is to identify gene-sets enriched with the signature genes using hypergeometric distribution. It was popular in the early times because it was simple and useful to infer the biological theme of signature genes. DAVID is a popular web-server that performs over-representation analysis³⁸. However, the biggest problem of this approach is to set the arbitrary cutoff for signature genes.

1.2.2.2 Functional class sorting

The cutoff-free method was devised to avoid setting such ambiguous cutoff for genes. Here, the gene-set score is directly evaluated by summarizing the gene-set member's scores obtained from omics data.

Besides, it is useful to detect gene-sets in which the individual genes show weak but consistent signals. Such pattern cannot be discovered using over-representation approach.

1.2.2.2.1 Gene-set enrichment analysis

One example pathway analysis method that implements the cutoff-free approach is the Gene-Set Enrichment Analysis (GSEA)³⁹. Since its paper was published in 2005, it has become the most widely used pathway analysis method. In GSEA, the input *a priori* gene-set scores (=enrichment score; ES) are evaluated using (weighted) Kolmogorov-Smirnov (K-S) statistic, which determines score based on the relative gene score rank distribution. For example, if members of a gene-set are distributed on the top ranks, it means that gene-set is up-regulated in overall. Similarly, member gene scores are concentrated in the bottom ranks, it represents the down-regulation of that gene-set. Detailed description for GSEA is in **2.3.1-Enrichment Score**.

Table 1.3. Popular pathway databases

Database	Pathway type	URL	Reference
Gene Ontology	<ul style="list-style-type: none"> Metabolic pathways Signalling pathways Protein-protein interaction 	http://www.geneontology.org	40
KEGG	<ul style="list-style-type: none"> Metabolic pathways 	http://www.genome.jp/kegg/	41
REACTOME	<ul style="list-style-type: none"> Metabolic pathways Signalling pathways 	http://www.reactome.org	42
RegulonDB	<ul style="list-style-type: none"> Transcription Factors / Gene Regulatory Networks 	http://regulondb.ccg.unam.mx/	43
PANTHER	<ul style="list-style-type: none"> Signalling pathways 	http://www.pantherdb.org	44
Ingenuity Pathway Analysis	<ul style="list-style-type: none"> Protein-Protein Interactions Metabolic Pathways Signaling Pathways Transcription Factors/ Gene Regulatory Networks Protein-Compound Interactions 	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/	45
NCI PID	<ul style="list-style-type: none"> Signaling Pathways 	http://pid.nci.nih.gov/	46
WikiPathways	<ul style="list-style-type: none"> Metabolic pathways Signalling pathways 	http://wikipathways.org/index.php/WikiPathways	47
Small Molecule Pathway DB	<ul style="list-style-type: none"> Metabolic pathways Signalling pathways 	http://www.smpdb.ca/	48
ConsensusPathDB	<ul style="list-style-type: none"> Protein-Protein Interactions Metabolic Pathways Signaling Pathways Transcription Factors / Gene Regulatory Networks Protein-Compound Interactions 	http://cpdb.molgen.mpg.de/CPDB	49
Pathway Commons	<ul style="list-style-type: none"> Protein-Protein Interactions Metabolic Pathways Signaling Pathways Protein-Compound Interactions 	http://www.pathwaycommons.org	50

1.2.2.3 Pathway topology-based method

In addition to over-representation analysis and functional class sorting, several methods based on pathway topology have been developed. Signaling Pathway Impact Analysis (SPIA) evaluates pathway significance by combining two p-values obtained from over-representation test and perturbation test⁵¹. CePa is a weighted gene-set analysis methods where the weights are determined by network centrality⁵². PathNet combines two types of evidences obtained from direct (p-value from DE analysis) and indirect evidence (inferred from pathway network neighborhood information) to get the signature genes. Then pathway significance is evaluated by hypergeometric test⁵³. Bayerlová et al. reported that these pathway topology-based methods showed better performance than classical enrichment-based methods under simulation setting with no overlapping gene-sets, but not in other settings⁵⁴. It means there are rooms to further develop this type of pathway analysis method (although not covered in this thesis...).

1.2.3 Competitive and self-contained gene-set analysis

Before performing gene-set analysis, we have to choose proper analysis method considering the null hypothesis. There are two methods mainly concerned: the competitive and self-contained methods. The null hypothesis (H_0) of each method is as follows:

- (1) H_0 of competitive method: Genes in a test gene-set are not more strongly associated with phenotype than the background genes.
- (2) H_0 of self-contained method: No genes in a test gene-set are associated with phenotype.

Thus, the competitive method tests the relative association of gene-sets compared to others. On the other hand, the self-contained method can significantly detect a gene-set if only few member genes are associated with the phenotype. Although it usually yields highly sensitive results, we should be careful in interpreting the result because gene-sets unrelated to phenotype can be specifically detected. Table 1.4 and 1.5 explains the gene-set analysis methods used for gene expression and GWAS data, respectively.

Table 1.4. Competitive and self-contained gene-set analysis methods for gene expression data.

Here, t_i and P_i are t-statistic and p-value of gene i , respectively, and m is gene-sets size.

Method	Statistic	Statistical test	Reference
Competitive methods			
Functional class score (FCS)	$FCS = \frac{\sum_{i=1}^m -\log(P_i)}{m}$	Gene permutation	55
Q1	$Q1 = \frac{\sum_{i=1}^m t_i}{m}$	Gene permutation	56
PAGE	$Z = \frac{\mu_G - \mu}{\delta/\sqrt{m}}$	Null distribution of $Z \sim N(0,1)$	57
Where μ and δ is average fold change and standard deviation of all genes, μ_G is average fold change of genes in test gene-set, and m is test gene-set size			
GSEA	Kolmogorov-Smirnov statistic	Sample or Gene permutation	39
CATEGORY	$Z = \frac{\sum_{i=1}^m t_i}{\sqrt{m}}$	Null distribution of $Z \sim N(0,1)$	58
GSA	$S_{max} = \max \left\{ \left \frac{\sum_{i=1}^m I(t_i > 0) \cdot t_i}{m} \right , \left \frac{\sum_{i=1}^m I(t_i < 0) \cdot t_i}{m} \right \right\}$	Sample permutation	59
Self-contained methods			
Globaltest	$Q = \frac{1}{m} \sum_{i=1}^m \frac{1}{\mu_2} [X'_i(Y - \mu)]^2$	Sample permutation or asymptotic distribution	60
Where X'_i is gene expression vector of gene i , Y is clinical outcome, μ is expectation of Y and μ_2 is the second central moment of Y under H_0 .			
FCS	$FCS = \frac{\sum_{i=1}^m -\log(P_i)}{m}$	Sample permutation	55
Q2	$Q2 = \frac{\sum_{i=1}^m t_i}{m}$	Sample permutation	56

Table 1.5. Competitive and self-contained gene-set analysis methods for GWAS summary data.

X_i is the gene score of i -th gene, μ and σ are mean and standard deviation of all gene scores, respectively, S is gene-set score, and m is gene-set size.

Method	Gene and/or gene-set statistic	Statistical test	Reference
Competitive methods			
GSA-SNP	$X_i = -\log_2(\text{k-th best SNP p-value})$ $S = \frac{\frac{1}{m} \sum_{i=1}^m X_i - \mu}{\sigma / \sqrt{m}}$	Null distribution of $Z \sim N(0,1)$ Restandardized GSA GSEA	61
MAGENTA	$X_i =$ Best SNP p-value corrected for confounding effects	Over-representation of top N-percentile of genes within each gene-sets is tested through comparison with random gene-sets.	62
INRICH	Genomic intervals associated phenotype are estimated first using PLINK LD clumping or tag SNP selection method. $S =$ the number of intervals (genes) overlapping with test gene-set.	The significance is evaluated through permutation process.	63
GOWINDA	Top N% SNPs are selected first	The significance of Over-representation with test gene-set is evaluated by permutations.	64
MAGMA	$X_i = \Phi^{-1}(1 - p_i)$ Where p_i is gene p-value estimated from mean or top χ^2 -statistic of SNPs within a gene.	From following linear model $Z_s = \beta_{0s} \vec{1} + S_s \beta_s + \varepsilon$ [$H_0: \beta_s = 0$] is tested. where β_s is the difference in association between gene-set members and background genes.	65
iGSEA4GWAS	$X_i = -\log_2(\text{best SNP p-value})$	SNP-permuting GSEA with significant proportion-based enrichment score ($SPES = ES \cdot k/K$), where k and K are proportion of significant genes (at least one SNP is included in the top 5% SNPs) of the gene-set and total gene list, respectively.	66
GSA-SNP2	$X_i = -\log_2(\text{k-th best SNP p-value})$ adjusted by SNP size $S = \frac{\frac{1}{m} \sum_{i=1}^m X_i - \mu}{\sigma^* / \sqrt{m}}$	Null distribution of $Z \sim N(0,1)$	67
Self-contained methods			
MAGMA	Same with MAGMA competitive method.	From following linear model $Z_s = \beta_0 \vec{1} + \varepsilon_s$ $H_0: \beta_0 = 0$ is tested.	65
sARTP	$w_{jk}^{(0)} = - \sum_{t=1}^{\min(q_j, c_k)} \log p_{j(t)}^{(0)}$	Direct simulation approach (DSA)	68

1.3 Biological network

Cells maintain life through consecutive biochemical reactions and interactions occurring among biomolecules such as metabolites, enzymes, transcription factors, signaling molecules and so on. Network is defined as a set of nodes and their relationships (edges). The interactions among biomolecules in a cell can be also represented as a complex network. There are five types of biological networks that are frequently used in bioinformatics as follows:

- 1) *Protein-protein interaction network*: The protein-protein interaction (PPI) represents the physical contact between proteins. PPIs occur in extensive cellular processes such as signal transduction, metabolism, electron transfer, transport across membranes, among others. Databases such as Database of Interacting Proteins (DIP), Biomolecular Interaction Network Database (BIND), Biological General Repository for Interaction Datasets (BioGRID), Human Protein Reference Database (HPRD), IntAct Molecular Interaction Database, Molecular Interactions Database (MINT), MIPS Protein Interaction Resource on Yeast (MIPS-MPact) and MIPS Mammalian Protein-Protein Interaction Database (MIPS-MPPI) provides validated PPI information,⁶⁹⁻⁷⁴. HIPPIE integrated these sources and provide reliable PPI information⁷⁵. STRING DB provides both known and predicted PPIs⁷⁶.
- 2) *Gene-regulatory network*: This network includes regulatory relationship between regulators (e.g., transcription factor, miRNA) and their target genes. Technologies such as ChIP-chip, ChIP-seq or Clip-seq are used to identify this network. ConsensusPathDB, Ingenuity Pathway Analysis⁷⁷ and Regulon DB provides this type of network.
- 3) *Gene co-expression network*: It represents the co-expression modules of genes in a specific cell condition. This network is generated from microarray or RNA-seq experiments followed by gene clustering analysis.
- 4) *Metabolic network*: It is the entire set of metabolic and physiological processes (e.g., fatty acid metabolism). Thus, it comprises the network of chemical compounds and enzymes involved in various biochemical reactions. KEGG, EcoCyc, and metaTIGER provides these networks⁷⁷⁻⁷⁸.
- 5) *Signaling network*: Cell signaling is a series of signal transduction that occurs within a cell or between cells to control the cellular action (e.g., PI3K/Akt signaling pathway). This process entails protein binding, phosphorylation, ubiquitination, acetylation and so on. The databases providing this network is represented in Table 1.3.

1.4 Research overview

Although many pathway analysis methods have been devised for gene expression or GWAS summary data, there have been still some limitations. First, most of the gene-set analysis methods for gene expression data had been designed for microarray. For RNA-seq data, seqGSEA⁷⁹, the use of log-transformed counts³¹ or pre-ranked GSEA with gene p-values from DE analysis had been suggested. However, there was a practical matter to apply these methods. That is, large number of RNA-seq data are composed the small number of samples due to the expensive sequencing cost. In this case, SeqGSEA, which implements sample permutation, is inappropriate to be used. Also, other two methods with gene permutation may yield many false positive results caused by inter-gene correlation among genes within same gene-set. In 2015, it was reported that the absolute statistic can effectively reduce the false positive rates in gene-permuting gene-set analysis of microarray data. In Chapter II, I tested whether GSEA with absolute gene statistic (absolute GSEA) exhibits same effect on RNA-seq data through simulation and real data analysis. For simulation test, a novel RNA-seq read count simulation method reflecting the inter-gene correlation was devised in this study. As a result, the absolute GSEA greatly improved the false positive control and overall discriminatory ability. The contents in this chapter are published in PLoS ONE in 2016 with the title ‘Improving Gene-Set Enrichment Analysis of RNA-Seq Data with Small Replicates.’⁸⁰

Next, I focused on the pathway analysis of GWAS summary data. Many of the competitive pathway analysis methods for GWAS summary data were too conservative to detect meaningful pathways. Some self-contained approaches were developed to increase the detection power, but there has been a concern that those methods may report gene-sets not relevant to phenotype as significant. In Chapter III, I will describe a powerful competitive gene-set analysis tools for GWAS summary data, named GSA-SNP2. By adjusting gene scores based on SNP size, it successfully increased the detection power while maintaining decent false positive control. The performance of GSA-SNP2 was validated using both simulation and real data. In addition, the GSA-SNP2 software provides gene network visualization within a gene-set or across significant gene-sets. The contents in this chapter are published in Nucleic Acids Research in 2018 with the title ‘Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2.’

Finally, Chapter IV describes a novel approach to infer cell condition-specific microRNA target network module by biclustering a large size of gene expression fold change profiles for a set of miRNA binding targets. The biclusters (network module) represent a set of miRNA binding motif-sharing genes commonly up-regulated (or down-regulated) under multiple cell conditions. The bicluster targets improved gain in certainty (sensitivity + specificity), and the net gain was further increased by incorporating functional network information. The analysis of cancer-related biclusters revealed that PI3K/Akt signaling pathway is strongly enriched in targets of a few miRNAs in breast cancer and

diffuse large B-cell lymphoma. Among them, five independent prognostic miRNAs were identified, and repressions of bicluster targets and pathway activity by mir-29 were experimentally validated. The BiMIR database provides useful search engine for biclusters of 459 human miRNAs.

Chapter II: Improving gene-set enrichment analysis of RNA-seq data with small replicates

2.1 Abstract

To identify deregulated biological pathways in a disease is important to understand the pathophysiology and find therapeutic targets of the disease. The gene-set enrichment analysis (GSEA) has been widely used for biological pathway analysis of microarray data, and it is also being applied to RNA-seq data. However, due to the high-sequencing cost, most RNA-seq data contain only small number of samples so far, which leads to perform gene-permuting GSEA method (or preranked GSEA). A critical problem of this method is that it yields many false positives results originated from the inter-gene correlation within gene-sets. I demonstrate that taking the absolute gene statistic in one-tailed GSEA greatly improves the false-positive control and the overall discriminatory ability of the gene-permuting GSEA methods for RNA-seq data. A novel simulation method to generate correlated read counts within a gene-set was devised to test performance, and a dozen of currently available RNA-seq enrichment analysis methods were compared, where the proposed methods outperformed others that do not account for the inter-gene correlation. Analysis of real RNA-seq data also supported the proposed methods in terms of false positive control, ranks of true positives and biological relevance. An efficient R package (AbsFilterGSEA) coded with C++ (Rcpp) is available from CRAN.

2.2 Introduction

The RNA-sequencing (RNA-seq) technology has facilitated a systematic analysis of the transcriptome in cells^{23, 81}. The biggest advantage of RNA-seq is much lowered background noise compared to the hybridization method (microarray). Thus, it has enabled more accurate quantification of gene expression level⁸². However, the differential expression (DE) analysis of RNA-seq data between two samples is not an easy task due to the different RNA composition and sequencing depth among samples as well as the discrete nature of RNA-seq data. Several between-normalization methods have been devised to make the gene expression levels among different samples comparable^{27, 83}, and a variety of methods have been developed to test the DE of each gene based on discrete probability models.^{28, 31, 33, 84-87}

The gene-set analysis has been used to interpret the DE analysis result. One approach is the GO analysis that estimates the over-representation of DE genes in a pre-defined gene-sets such as Gene Ontology (GO) terms.⁸⁸⁻⁸⁹ The gene-set enrichment analysis (GSEA) is another useful approach³⁹. Unlike GO

analysis, it does not use the cutoff threshold to identify the DE genes. Instead, it utilizes the (weighted) Kolmogorov-Smirnov (K-S) statistic to test whether genes contributing to the phenotype are ‘enriched’ in each gene-set. Therefore, GSEA can detect the subtle but coordinated changes in a gene-set and has been widely used to find important pathways or functions in various diseases and cell conditions from microarray data ^{61, 90-92}.

The pathway analysis methods and tools for RNA-seq have recently been devised based on methods designed for microarray ^{31, 93-95}. One of the issues in applying GSEA to RNA-seq data is the normalization of read count data. Voom method transforms the read counts into microarray-like data for which most linear-model based methods developed for microarray can be applied ³¹. GSAAseqSP tool ⁹⁴ adopted TMM or DESeq normalization methods ^{27, 85} which are able to address both the different depths and RNA compositions between samples. Another important thing to consider is the small sample sizes in RNA-seq data. Although the sequencing cost has been lowered so fast, it is still expensive. Thus, most laboratories have no choice but to produce only a few replicates for each condition ⁹⁶. The sample-permuting GSEA (GSEA-SP) is inappropriate to apply to such small replicate data. Instead, the gene-permuting GSEA (GSEA-GP) is used in this case. However, the GSEA-GP generates a lot of false positive gene-sets due to the inter-gene correlation in the gene expression.

In this study, it was demonstrated that the *absolute gene statistic* improved the false positive control and overall discriminatory ability of GSEA-GP of RNA-seq data. Although the property was shown in microarray data ⁹⁷, it was not tested in RNA-seq data yet. The RNA-seq read counts were modeled and simulated using discrete probability (negative binomial distribution) ^{84, 98}, and a simulation method to generate ‘correlated’ read counts within a gene-set was newly devised to compare the performance of GSEA methods for RNA-seq data. Note that the inter-gene correlation has a critical effect on the performance of gene-set level analysis, but has not been considered so far for the counting data because of the lack of such a simulation method.

Here, the one-tailed GSEA which takes the maximum positive deviation of the K-S statistic as a gene-set enrichment score was used for more precise gene-set analysis. Based on this result, I also propose filtering the GSEA-GP results with those obtained from the absolute GSEA-GP to effectively reduce false positives. The performances of the absolute GSEA and its filtering method were demonstrated for simulated and real RNA-seq data.

2.3 Materials and Methods

2.3.1 Absolute gene-permuting GSEA and filtering

In many cases, the replicate size is too small in RNA-seq data to carry out GSEA-SP (e.g., $n < 10$). In that case, the GSEA-GP is used instead. However, it produces a lot of false positive results because of the inter-gene correlation within gene-set ⁹⁹⁻¹⁰³. Recently, it has been shown that incorporating the

absolute gene statistic in GSEA-GP considerably reduces the false positive rate and improves the overall discriminatory ability in analyzing microarray data⁹⁷. Therefore, it was tested whether the absolute statistic shows a similar effect in RNA-seq data analysis. In addition to replacing the gene statistic with their absolute values¹⁰⁴, the absolute GSEA was modified as a one-tailed test in this study by considering only the ‘positive’ deviation in the K-S statistic. There are two reasons for this adjustment. First, simple substitution of gene scores by taking absolute values in GSEA can produce a small number of ‘down-regulated’ gene-sets which are meaningless in an absolute enrichment analysis. Second, it gives more precise null distribution of gene-set statistic: In the original GSEA algorithm, the maximum positive and negative deviation values are compared and only the larger absolute value between the two is selected for the gene-set score. This means the minor maximum deviation values are all excluded in constituting the gene-set null distributions. By taking only the positive deviation values, every gene-set contributes to the null distribution.

Gene scores: Four gene scores were considered for normalized read as follows:

- (1) **Moderated t-statistic (mod-t):** A modified two-sample t-statistic

$$\tilde{t}_i = \frac{\mu_i^1 - \mu_i^2}{\tilde{s}_i \sqrt{v_i}}$$

where μ_i^n is the mean read count of i th gene, g_i in class n , and \tilde{s}_i is a shrinkage estimation of the standard deviation of g_i . This statistic is useful for small replicate data and is implemented using the limma R package¹⁰⁵⁻¹⁰⁶

- (2) **Signal-to-Noise ratio (SNR):** The SNR (S_i) is calculated as

$$S_i = \frac{\mu_i^1 - \mu_i^2}{\sigma_i^1 + \sigma_i^2}$$

where σ_i^n is the standard deviation of expression values of g_i in class n .

- (3) **Zero-centered rank sum (Ranksum):** This two-sample Wilcoxon statistic is introduced by Li and Tibshirani³³. For g_i , the rank sum test statistic (T_i) is calculated as,

$$T_i = \sum_{j \in C_1} R_{ij} - \frac{n_1 \cdot (n + 1)}{2}$$

where R_{ij} is the rank of expression level of j^{th} sample among all counts of g_i , C_1 is a set of sample indexes in the first phenotypic class, n_1 is the sample size of C_1 and n is the total sample size. Note that $E(T_i) = 0$.

- (4) **Log fold-change (logFC):** Log fold-change ($\log FC_i$) for g_i is calculated as

$$\log FC_i = \log_2 \frac{\mu_i^1}{\mu_i^2}$$

Absolute GSEA: GSEA algorithm identifies functional gene-sets showing a coordinated gene expression change between case and control samples from gene expression profiles. Given gene scores, GSEA implements a (weighted) K-S statistic to calculate the enrichment score (ES) of each pre-defined gene-set.

(1) Enrichment score

Let S be a gene-set and r_i be the gene score of g_i . Then, the enrichment score $ES(S)$ is defined as the maximum deviation of $p_{hit} - p_{miss}$ from zero, that is

$$ES(S) = \begin{cases} \max_i(p_{hit,i} - p_{miss,i}), & \text{if } \left| \max_i(p_{hit,i} - p_{miss,i}) \right| \geq \left| \min_i(p_{hit,i} - p_{miss,i}) \right| \\ \min_i(p_{hit,i} - p_{miss,i}), & \text{if } \left| \max_i(p_{hit,i} - p_{miss,i}) \right| < \left| \min_i(p_{hit,i} - p_{miss,i}) \right| \end{cases}$$

where

$$p_{hit,i} = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^q}{N_R}, \quad p_{miss,i} = \sum_{\substack{g_j \in S^c \\ j \leq i}} \frac{1}{(N - N_H)}, \quad N_R = \sum_{g_j \in S} |r_j|^q$$

N is the total number of genes in the dataset, N_H is the number of genes included in S and q is a weighting exponent which is set as one in this study as recommended³⁹. (For the classical K-S statistic, $q = 0$)

(2) ES for one-tailed absolute GSEA

The absolute GSEA is simply performed by substituting the gene scores by their absolute values, but the ranks of gene scores are quite different from the original GSEA algorithm in calculating the K-S statistic. For the one-tailed test, only the positive deviation $ES(S) = \max_i(p_{hit,i} - p_{miss,i})$ is considered for the gene-set score.

Then, the gene permutations are applied, and the corresponding ES's are calculated and normalized for evaluating the false discovery rate of each gene-set³⁹.

Filtering with absolute GSEA: To decrease the false positive results in the GSEA-GP, it is recommended to use the absolute GSEA-GP results for filtering false positives from the ordinary GSEA-GP results. In other words, only the gene-sets that are significant in both ordinary and the one-tailed absolute GSEA are considered significant. In this way, more reliable gene-sets with directionality can be obtained. In all the analyses presented in this paper, the same FDR cutoff is applied for both ordinary and absolute methods, but different cutoffs can also be considered for stricter or looser filtering.

2.3.2 Simulation of the read count data with the inter-gene correlation

High inter-gene correlation within gene-sets severely increases the false positive rates in gene-permuting gene-set analysis methods (a.k.a. competitive analysis)^{99, 103}. The inter-gene correlation of microarray data can be modelled using multivariate normal distribution^{97, 101, 107}, but it cannot be

directly applied for ‘discrete’ read count data. Here, a novel simulation method for read count data with inter-gene correlation in each gene-set is described. $N=10,000$ genes are considered and the replicate sizes for the test and control groups are n_1 and n_2 , respectively.

Step 1. Parameter estimation and read count generation: The read count X_{ij} of i th gene in j th sample has been modeled by an over-dispersed *Poisson* distribution, called negative binomial (NB) distribution^{84-85, 98} denoted by $X_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$ where μ_{ij} and $\sigma_{ij}^2 = \mu_{ij} + \varphi_i \mu_{ij}^2$ are the mean and variance, respectively, and $\varphi_i \geq 0$ is the dispersion coefficient for gene g_i . Here, $\mu_{ij} = s_j \mu_i$, where s_j is the ‘size factor’ or ‘scaling factor’ of sample j and μ_i is the expression level of g_i . In this simulation, all size factors s_j were set as 1 for simplicity. The mean and gene-wise dispersion parameters of 10,000 genes (average read count >10) were estimated from TCGA kidney cancer RNA-seq data (denoted as TCGA KIRC)¹⁰⁸. The edgeR R package was used to estimate both parameters⁹⁸. The read counts were generated using the R function ‘rnbinom’ where the inverse of the estimated dispersion φ_i was input as the ‘size’ argument. This method generates read counts that are independent between genes.

Step 2. Generation of read count data with the inter-gene correlation: Given a gene-set S with K genes, the inter-gene correlation can be generated by incorporating a common variable within the gene-set. Let μ_i and $\varphi_i, i=1,2,\dots,K$ be the mean and tag-wise dispersion of g_i in the gene-set and C_{ij} be the read count generated from these parameters (Step 1). Let $P_S = \{p_1, p_2, \dots, p_{n_1+n_2}\}$ be probability values randomly sampled from the uniform distribution $U(0,1)$. Then, for each g_i , the probability values in P_S are converted to a read count $C_{ij}^*, j=1,2,\dots, n_1+n_2$ using the inverse function of the individual gene’s distribution $X_i \sim NB(\mu_i, \varphi_i)$ such that $p_j \approx P(X_i \leq C_{ij}^*)$. In short, C_{ij}^* are generated from the common uniform distribution via the gene-wise NB distribution. The ‘correlated’ read count for i th gene in j th sample is then obtained by the weighted sum of the original count C_{ij} and the ‘commonly generated’ count C_{ij}^* as follows:

$$M_{ij} = \lceil (1 - \alpha) \cdot C_{ij} + \alpha \cdot C_{ij}^* \rceil$$

where $\alpha \in [0,1]$ is the mixing coefficient that determines the strength of the inter-gene correlation and $\lceil \cdot \rceil$ rounds the value to the nearest integer. One problem with this count is that its variance is reduced as much as $(2\alpha^2 - 2\alpha + 1)$ because

$$\text{Var}(M_{ij}) \approx (1 - \alpha)^2 \cdot V(C_{ij}) + \alpha^2 \cdot V(C_{ij}^*) = (2\alpha^2 - 2\alpha + 1) \cdot \sigma_{ij}^2$$

To remove this factor, an inflated dispersion φ_i' was used derived from the equation

$$(2\alpha^2 - 2\alpha + 1) \cdot (\mu_i + \varphi_i' \mu_i^2) = \mu_i + \varphi_i \mu_i^2$$

$$\varphi_i' = \frac{1 + \varphi_i \mu_i}{\mu_i \cdot (2\alpha^2 - 2\alpha + 1)} - \frac{1}{\mu_i}$$

instead of φ_i in generating C_{ij} and C_{ij}^* . The relationship between α and inter-gene correlation is shown in Figure 2.1.

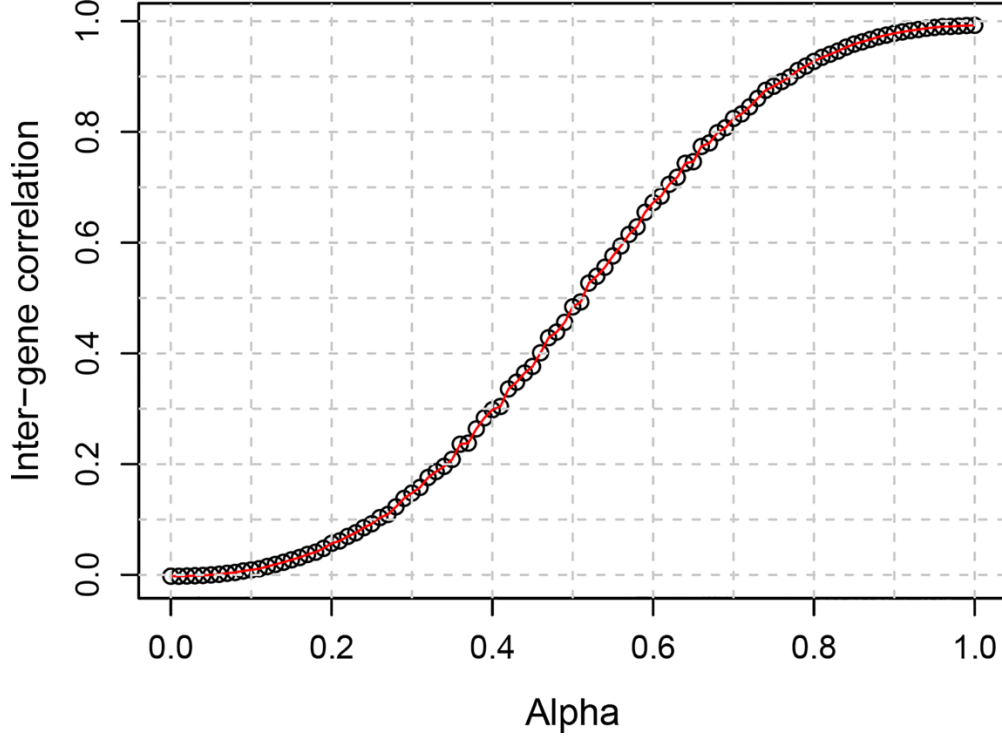


Figure 2.1. The relationship between the mixing coefficient (alpha) and the average inter-gene correlation.

2.3.3 Biological relevance measure of a gene-set

To measure the functional relevance of gene-sets filtered by absolute GSEA, a gene-set score based on literatures (PubMed abstract) was designed. Here, it was assumed that the significantly altered gene-sets contain genes playing important roles in the alteration of corresponding cell (tissue) condition. For a significant gene-set S , its relevance with a specific tissue T is scored by the log geometric average of the abstract counts as follows:

$$L(S) = \frac{1}{K} \sum_{i=1}^K \log(A_{T,i}) \quad (1)$$

where K is the gene-set size and $A_{T,i}$ is the number of PubMed abstracts where both the keywords related to the tissue T and the name of g_i co-occur. The literature mining was conducted using RISmed R package¹⁰⁹.

2.3.4 RNA-seq data handling and gene-set condition

The RNA-seq raw read counts were normalized by the DESeq⁸⁵. To make the logFC stable for small read counts, the lower 5% of normalized counts larger than zero were added to the normalized counts. Such pseudocount does not change other types of gene scores.

2.3.5 Gene-set size

The ‘gene-set size’ represents the number of overlapping genes between the original pathway and input RNA-seq dataset. In this study, the gene-set size was constrained to 10~300.

2.3.6 AbsfilterGSEA R package

I developed a CRAN R package ‘AbsFilterGSEA’ that performs both original and absolute gene-permuting GSEA¹¹⁰. Here, the input raw read count matrix is normalized by DESeq method⁸⁵. It also accepts an already normalized dataset. It is quite fast because the GSEA part was implemented with C++. The integration of C++ code to the R package was done by Rcpp package¹¹¹.

2.4 Results

2.4.1 Comparison of gene-permuting GSEA methods for simulated read count data

The performance of twelve GSEA-GP methods for small replicate data were compared using simulation dataset reflecting the inter-gene correlation within gene-sets (See Section 2.3.2). The simulated read count data included 10,000 genes and 100 non-overlapping gene-sets each of which contained 100 genes.

First, the false positive rates ($FDR < 0.1$) of the GSEA-GP methods for the four gene statistics (mod- t , SNR, Ranksum and FC) and their absolute counterparts were measured using the simulated read count datasets with four different levels of inter-gene correlation, LOW (0~0.05), 0.1, 0.3 and 0.6 within each gene-set. Two, three and five replicates in each sample group were tested and no DE genes were included. This test was repeated twenty times and their average false positive rates were depicted in Figure 2.2A and 2.2D for three and five replicates, respectively. Figure S2.1A shows the result for two-replicate case. A recently developed competitive method, Camera combined with the voom normalization^{31, 101}, the bias-adjusted random-set method (RNA-Enrich)⁹⁵ as well as two preranked GSEA methods³⁹ were also compared. The preranked GSEA (unweighted) was implemented using the GSEA R-code³⁹ where the ranks of genes were determined according to either the p -values resulted from the differential expression analysis using edgeR⁹⁸ package or the simple absolute fold-changes of the normalized count data. Note that SeqGSEA⁷⁹ provides only sample-permuting GSEA which is not useful for small replicates, and GSAAseqSP⁹⁴ provides a gene-permuting GSEA method which is

almost same as GSEA-GP described in this paper (I checked they yielded nearly the same results for the simulated count data). Although it is described that GSAAseqSP uses the absolute gene scores, they are only used for the step-sizes in K-S statistic, and it is far from the ‘absolute’ enrichment analysis. The false positive rates of GSEA-GP for the four ordinary gene statistics and the two preranked methods showed upsurge with the increasing inter-gene correlation. However, the increase rates of false positive rates for the four absolute GSEA methods were considerably lower than those for the ordinary statistics.

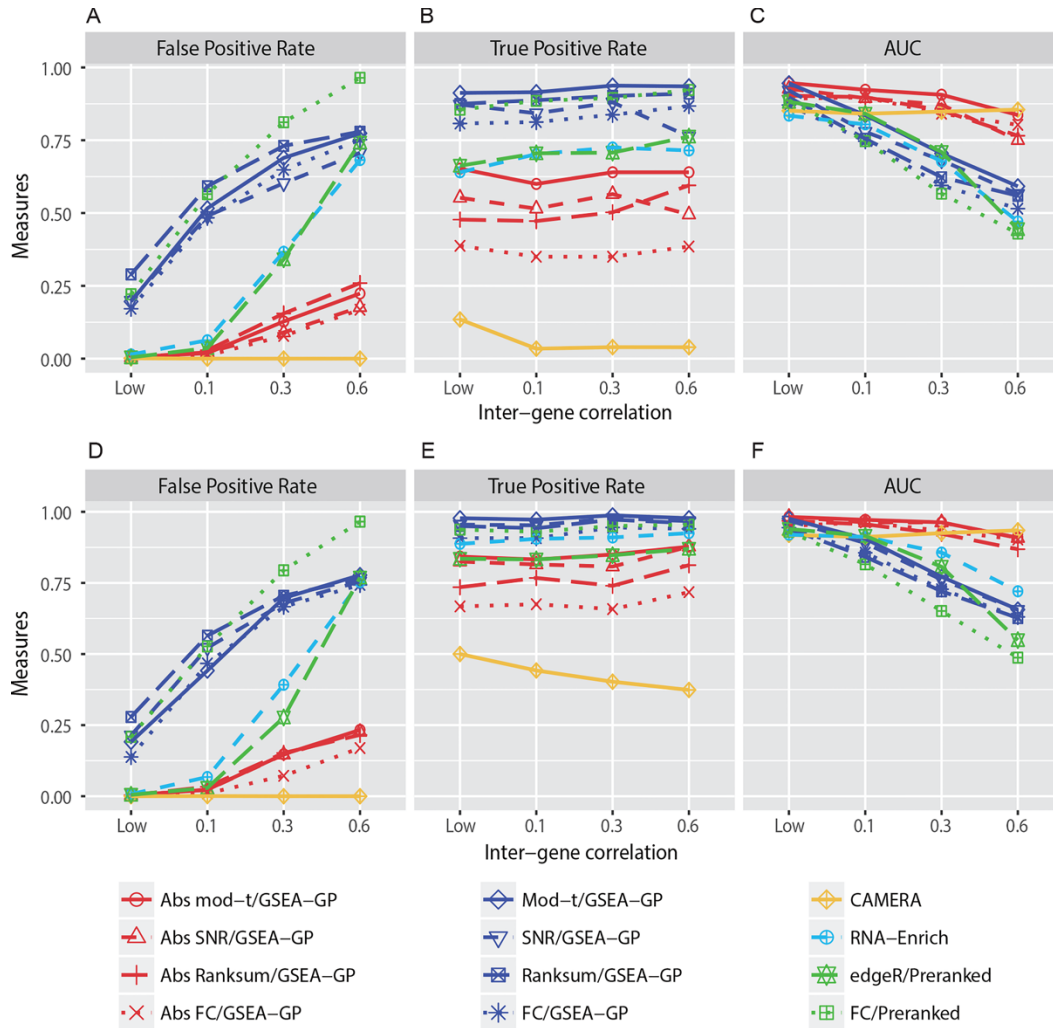


Figure 2.2. Performance comparison of gene-permuting GSEA methods for simulated read counts.

GSEA-GP methods combined with eight gene statistics, (moderated t-statistic, SNR, Ranksum, logFC and their absolute versions), Camera combined with voom normalization, RNA-Enrich and two preranked GSEA methods for edgeR p-values and FCs were compared for false positive rate, true

positive rate and area under the receiver operating curve using simulated read count data with three (A-C) and five replicates (D-F).

For example, when three replicates were used, even for a moderate inter-gene correlation 0.1, the false positive rates for the original statistics were approximately 50% or higher while only a few false positive sets were detected for the absolute methods (1 ~ 3%). Camera yielded no false positives for each correlation level. Overall similar trends were observed with five replicates, but the absolute *mod-t* and absolute SNR exhibited nearly the same AUCs. RNA-Enrich and the edgeR/preranked methods exhibited relatively better false positive rates compared to the GSEA-GP and FC/Preranked methods. Next, 20% of the gene-sets (20 gene-sets) in the data generated above were replaced with differentially expressed gene-sets to compare the power (true positive rate) and the overall discriminatory abilities (ROC). These gene-sets included 20~80% (uniformly at random in each gene-set) of DE genes whose mean counts in the test or control group were multiplied by 1.5~2.0 with which the read counts in the corresponding group were regenerated. In DE gene-sets, weak inter-gene correlations (0~0.05) were randomly assigned while the non-DE gene-sets were assigned with four different inter-gene correlation levels. The corresponding powers and the area under the ROC curves (AUCs) were then obtained for the twelve methods compared (Fig. 2.2B, 2.2C, 2.2E and 2.2F). The preranked GSEA with FCs and GSEA-GP methods had the highest levels of power, but their AUCs rapidly declined as the inter-gene correlation level was increased because of their poor false positive controls. With the inter-gene correlation of 0.6, their performances were close to a random prediction ($AUC \approx 0.5$). On the other hand, the absolute GSEA-GP methods and Camera exhibited stable and good AUCs irrespective of the inter-gene correlation level. The ROC curves (average of 20 repetitions) of the twelve gene-permuting GSEA methods for the inter-gene correlation 0.3 are illustrated in Figure 2.3. For the two-replicate data, the false positive rates were similar to those of triplicate case, but the powers and AUCs were rather lowered (Fig. S2.1a). While the *mod-t* still exhibited best powers and AUCs among the absolute methods, the power of SNR was considerably lowered, which necessitates the moderated gene statistic in GSEA of small replicate data. Lastly, different inter-gene correlations were randomly assigned for gene-sets in a dataset, and two, three and five replicate cases were tested (Fig. S2.1b-d). The absolute *mod-t* still exhibited best AUCs in most cases and exhibited overall similar trends as the identical correlation cases.

Overall, these results indicate that the absolute GSEA-GP provides an excellent false positive control and improves the overall discriminatory ability of GSEA-GP. Although the ordinary GSEA-GP methods exhibited best powers, they suffered from prohibitively high false positive rates resulting in very poor ranks of true positives (AUCs). Compared with Camera, the absolute methods yielded a little more false positives, but exhibited better power and overall discriminatory ability (correlation < 0.6). In

general, for small replicate datasets, not all of the true positives may be identified perfectly by any method, but it would be important to discern some of the truly altered gene-sets reliably.

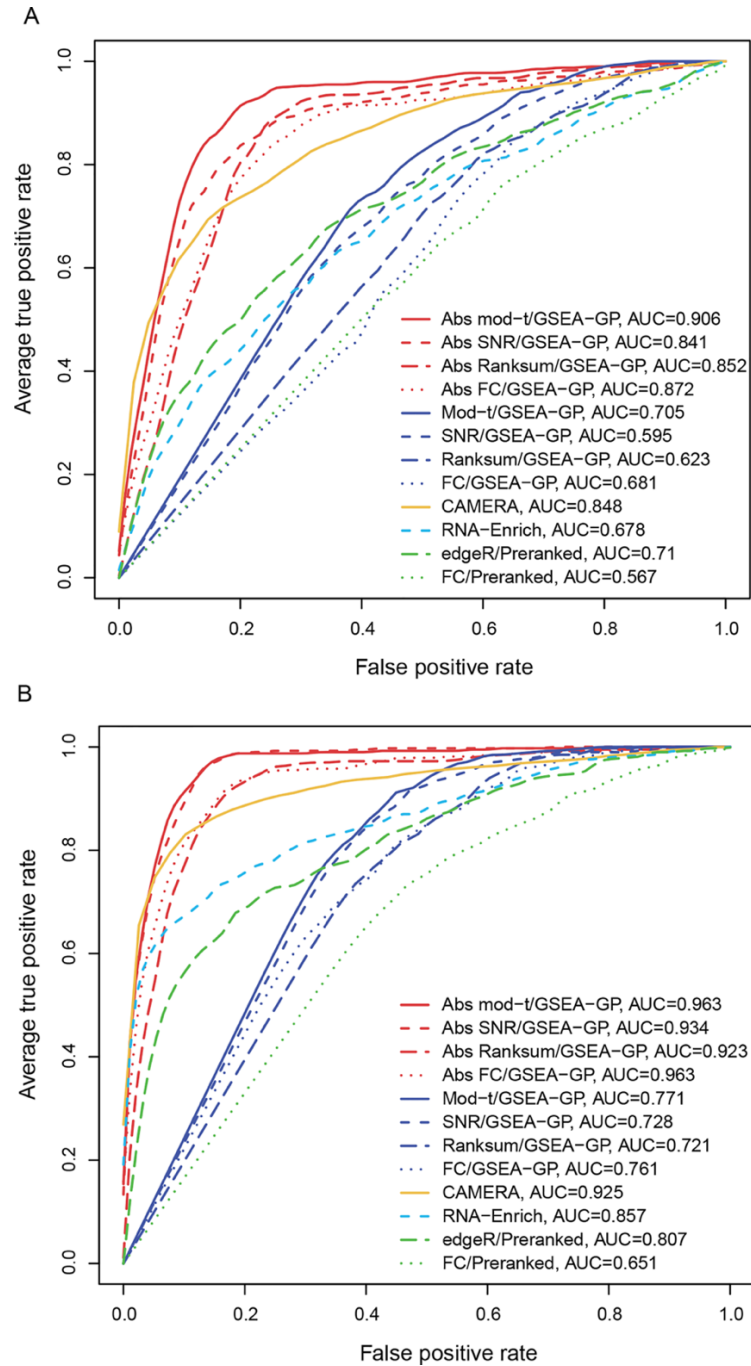


Figure 2.3. Average receiver operating characteristic (ROC) curves.

The average ROC curves (20 repetitions) of the twelve gene-permuting GSEA methods applied to simulation data with the inter-gene correlation of 0.3 for (A) three and (B) five replicate cases

2.4.2 Comparison of GSEA methods for RNA-seq data

The performances of GSEA methods were compared for published RNA-seq datasets in several aspects. First, two RNA-seq datasets denoted by Pickrell and Li data, respectively, were analyzed for comparing power and accuracy as follows:

The Pickrell data were generated from the lymphoblastoid cell lines of 69 unrelated Nigerian individuals (29 male and 40 female)¹¹². To analyze the chromosomal differences in expression between male and female, MSigDB C1 (cytogenetic band gene-sets)¹¹³⁻¹¹⁵ was used for analysis. The GSEA-SP with SNR gene score was applied for the total dataset which resulted in two significant gene-sets ‘chryq11’ (FDR=0.00143) and ‘chrxp22’ (FDR=0.0514) both of which were sex-specific. These two gene-sets were significantly up-regulated in male and female groups, respectively. Since the GSEA-SP controls the false positives well, these two gene-sets were regarded as true positives. Then, five samples were randomly selected from each group to constitute a small replicate dataset and GSEA-GP methods with or without absolute filtering, Camera, edgeR/Preranked methods were compared for this small replicate dataset. This process was repeated ten times. Using $\text{mod-}t$ and $\log\text{FC}$ as the gene scores, on average, the GSEA-GP yielded 33.9 and 19.9 significant (FDR<0.25) gene-sets including 1.5 and 1.1 true positives, respectively. On the other hand, GSEA-GP with the absolute filtering resulted in only 3.67 and 2.9 significant gene-sets which included 1.11 and 1 true positives for the $\text{mod-}t$ and $\log\text{FC}$ gene scores, respectively. For these five-replicate datasets, Camera did not detect any significant gene-set, and the edgeR/Preranked detected as many as 137.4 which included 1.8 true positives. This result implies that the absolute filtering method effectively reduces the false positives resulted from GSEA-GP while maintaining a good statistical power.

A similar trend was observed with the Li dataset. The Li data¹¹⁶ were generated from LNCaP cell lines with three samples treated with dihydrotestosterone (DHT) and four control samples. The MSigDB C2 (curated gene-set) was used for analysis and the six gene-sets containing the term ‘androgen’ were regarded as potential true positives since DHT is a kind of androgen, though there can be other truly altered gene-sets. When the GSEA-SP with $\text{mod-}t$ and $\log\text{FC}$ gene score was applied for this small replicate dataset, as expected, only one and no ‘androgen’ gene-set was significant (FDR<0.1), respectively. On the other hand, GSEA-GP with $\text{mod-}t$ and $\log\text{FC}$ gene scores yielded as many as 187 and 569 significant gene-sets, respectively, which included four ‘androgen’ gene-sets with FDR≤0.0067. When the absolute filtering was applied, the numbers of significant gene-sets were dramatically reduced to *eight* (Table 2.1) and 242, which included three and four ‘androgen’ gene-sets, respectively. Of note, the top three gene-sets were ‘androgen’ terms for the $\text{mod-}t$ score. The absolute GSEA filtering with SNR score provided a similar result. Camera detected only two ‘androgen’ gene-sets within 101 significant gene-sets with FDR=0.00836 and 0.0195, respectively. RNA-Enrich and edgeR/Preranked

were so sensitive for this dataset that 1108 and 782 sets were significant (FDR<0.1). RNA-Enrich and edgeR/Preranked detected *four* and three androgen terms within top 52 and 91 gene-sets.

Overall, the results for real data analysis were concordant with the simulation results. GSEA-GP yielded a large number of significant gene-sets most of which seemed to be false positives. The absolute filtering method considerably reduced false positives at the cost of only small loss of power. Camera exhibited a strict false positive control, but its power was relatively weak. In particular, the absolute filtering with mod-*t* score exhibited a high precision and a good power in both datasets.

Table 2.1. Significant gene-sets detected by the absolute GSEA-GP filtering (FDR<0.1) with the mod-*t* score (DHT-treated and control LNCaP cell line).

Gene-set name	FDR	Literature score
Response to androgen (down, Nelson)	0	2.15
Response to androgen (up, Nelson)	0	1.87
Response to androgen (up, Wang)	1.63×10^{-4}	1.37
PKD1 targets (up, Piontek)	2.79×10^{-4}	1.78
Reactome Amino acid synthesis and interconversion transamination	2.27×10^{-2}	1.94
Response to forskolin (up, Wang)	3.02×10^{-2}	1.42
AML cluster 11 (Valk)	5.12×10^{-2}	1.17
Breast basal vs. luminal (up, Huper)	4.40×10^{-2}	1.53

2.4.3 Effects of the absolute filtering on false positive control and biological relevance

Here, the effects of the absolute filtering were analyzed for real data in two other aspects. The first one is the false positive rate as investigated with the variance inflation factor (VIF). The false positive rate of a competitive gene-set analysis method is known to be determined by VIF which is defined as:

$$\text{Var}(\text{gene set statistic}) = \text{Var}_{\text{i.i.d.}}(\text{gene set statistic}) \times \text{VIF}$$

where $\text{Var}_{\text{i.i.d.}}$ is the variance of a gene-set statistic under the assumption that genes in each gene-set have independent expression values. For a linear gene-set statistic, the VIF is explicitly represented as a function of the gene-set size (K) and the average inter-gene correlation ($\bar{\rho}$)^{101, 117} as follows:

$$\text{VIF} = 1 + (K - 1)\bar{\rho} \quad (2)$$

To compare the false positive rates of the GSEA-GP and the absolute GSEA-GP methods approximately, VIF distributions (2) of the significant gene-sets were compared for two TCGA RNA-seq datasets (KIRC kidney tumor vs. normal¹¹⁸ and BRCA breast tumor vs. normal¹⁰⁸). These datasets were comprised of a large number of cancer and normal samples (144 and 216 for the KIRC and BRCA,

respectively) with which the average inter-gene correlation may be accurately estimated. In each dataset, five cancer samples and five normal samples were randomly drawn to constitute a small replicate dataset, to which GSEA-GP was applied using the gene scores logFC and absolute logFC, respectively. Then, the VIFs were compared between two classes of significant gene-sets. One is the gene-sets that are significant only in the ordinary GSEA-GP (class A) and the other is those that are significant in both the ordinary and absolute GSEA-GP methods (class B). Note that the total samples in each dataset were used to calculate $\bar{\rho}$. This process was repeated ten times and the corresponding VIF distributions were compared. In most cases, VIFs of class B were significantly smaller than those for class A. For KIRC data, all the ten sub-datasets exhibited significantly smaller VIFs in class B (Wilcoxon ranksum p-value<0.05; smallest p-value 6.15E-8). Similarly, seven out of ten sub-datasets derived from BRCA data showed significance (smallest p-value 2.14E-5). This indicates the absolute filtering method substantially reduces the false positives in real data analysis. The second aspect is the tissue-specific relevance score (1). As the above case, five samples were randomly selected from each group, and the literature relevance scores between the class A and B sets were compared for both KIRC and BRCA datasets. As a result, nine and four out of ten sub-datasets, the relevance scores in class B were significantly larger for the KIRC and BRCA datasets, respectively (smallest p-values: 7.87E-12 and 1.06E-5, respectively). These results indicate that the absolute filtering results in highly reliable and biologically relevant gene-sets.

2.5 Discussion

Since the advent of RNA-seq technology until recently, various methods to identify DE genes from the RNA-seq read count data have been developed^{31, 84, 98, 116}. One notable feature shared by DE analysis methods is that they yield quite a number of DE genes. Typically, hundreds to thousands genes are differentially expressed with RNA-seq data of two sample groups. RNA-seq is known to provide a much improved resolution in quantitating gene expression compared to that of microarray⁸¹, which may have increased the sensitivity of DE analysis for RNA-seq data.

With the increased resolution and sensitivity, the pathway analysis or GSEA are expected to play a crucial role in genomic studies with their ability to detect the ‘subtle but coordinated’ changes in a gene-set³⁹. However, in many cases, only GO analysis has been applied for interpreting RNA-seq data¹¹⁹. The low application rate of pathway analysis or GSEA for RNA-seq may be ascribed to the lack of tools that are specifically designed for RNA-seq data. The popularly used GSEA software³⁹ developed for microarray analysis can be used for RNA-seq data by normalizing the read count data ‘appropriately’ or simply applying the gene-permuting method (preranked GSEA) after ranking the gene differential scores using another software (e.g. edgeR or DESeq).

Since the majority of RNA-seq experiments have generated only small replicates, the preranked GSEA methods were often used for function and pathway analysis. However, gene-permuting methods usually result in a great number of false positives due to the inter-gene correlation whatever the replicate sizes are. To date, Camera¹⁰¹ has been the only method to control the false positive gene-sets caused by the inter-gene correlation in analyzing small replicate read count data, but its statistical power was quite weak. In this study, I showed one-tailed absolute GSEA manifests an excellent false positive control and a good statistical power for analyzing small replicate RNA-seq data.

To compare the performance of GSEA methods, read count data incorporating the inter-gene correlation were newly simulated. It is crucial to consider the inter-gene correlation in evaluating gene-set analysis methods. The analysis results for the simulated and RNA-seq data commonly demonstrated the effectiveness of the suggested method. As such, the method and tool presented in this study may facilitate the pathway analysis of RNA-seq data with small replicates.

2.6 Supplementary information of Chapter II

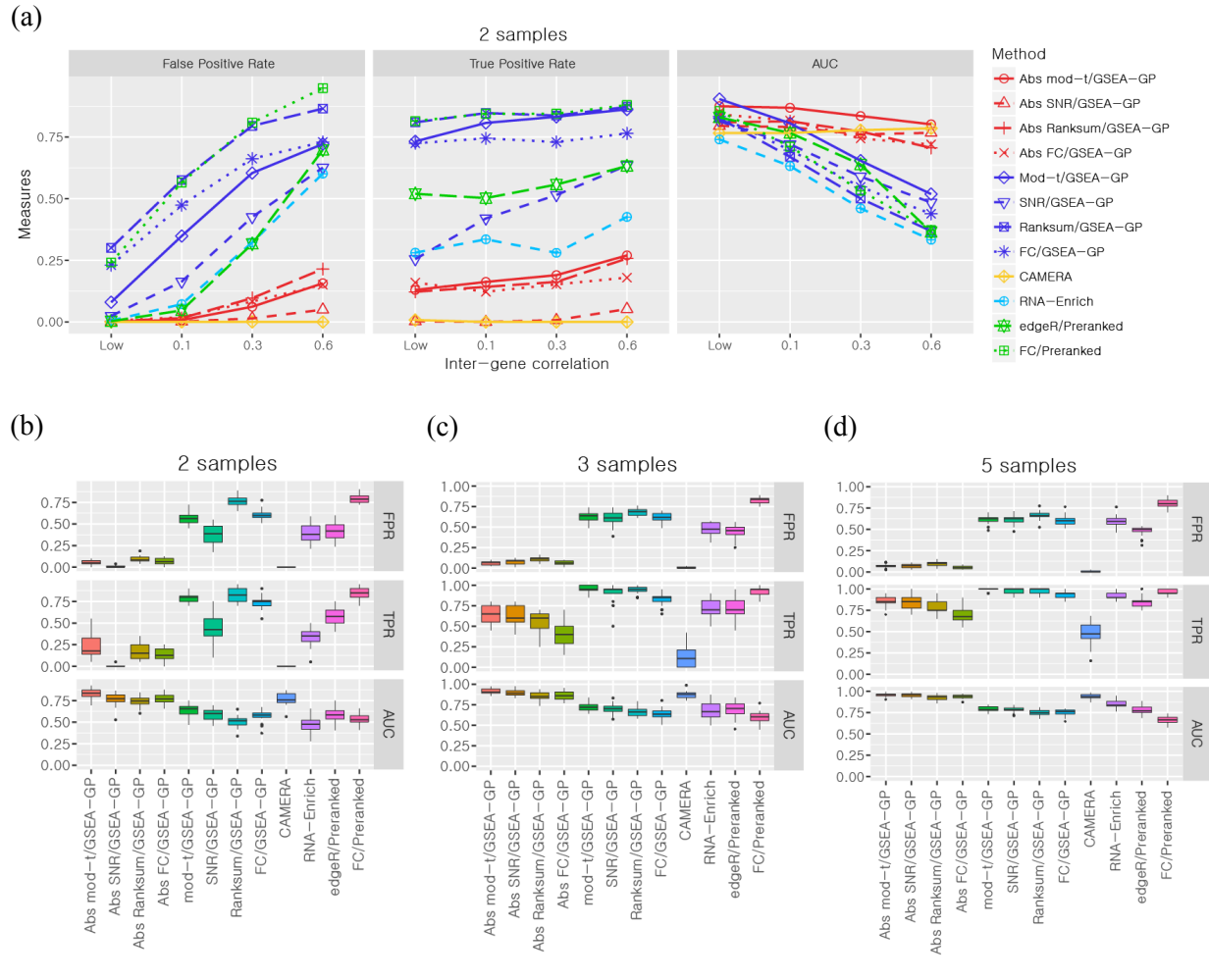


Figure S2.1. Performance comparison of gene-permuting GSEA methods for simulated read counts.

GSEA-GP methods combined with eight gene statistics, (moderated t-statistic, SNR, Ranksum, logFC and their absolute versions), Camera combined with voom quantile normalization, RNA-Enrich and two preranked GSEA methods for edgeR p-values and FCs were compared for false positive rate, true positive rate and area under receiver operating curve (a) by increasing the inter-gene correlation of simulated read count data composed of two replicates, (b) or assigning various random inter-gene correlations (0~0.6) to each simulation dataset composed of two, (c) three, (d) and five replicates

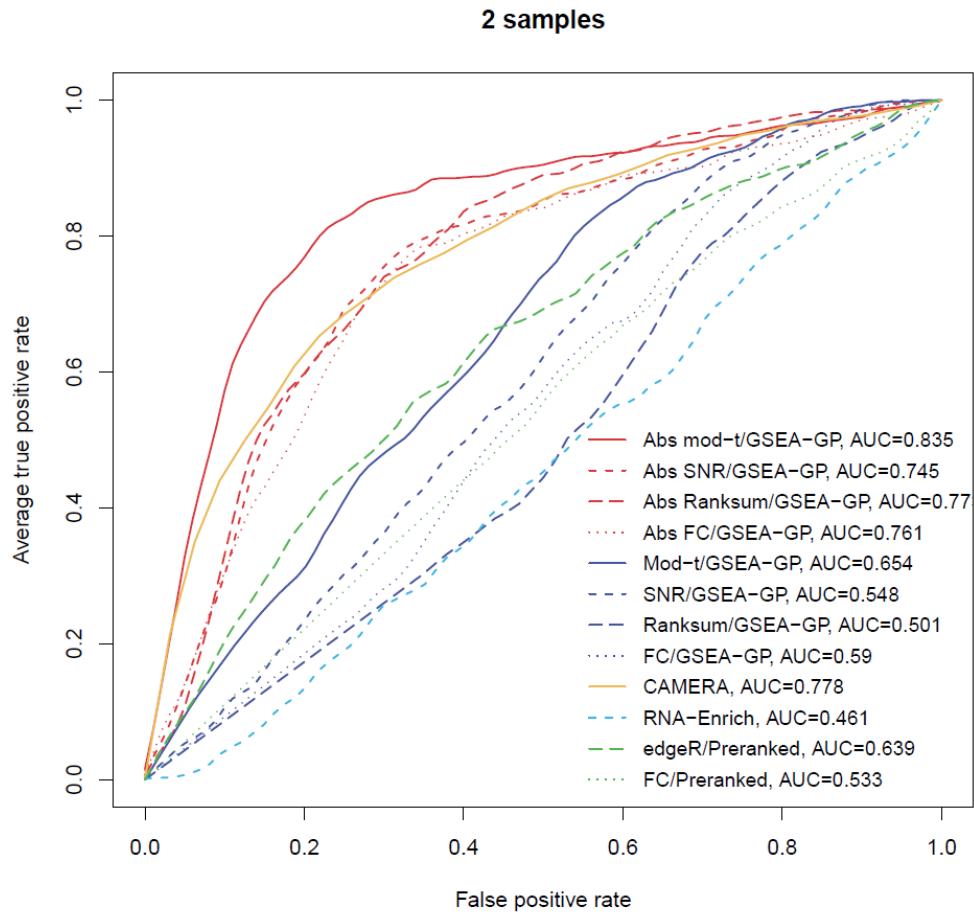
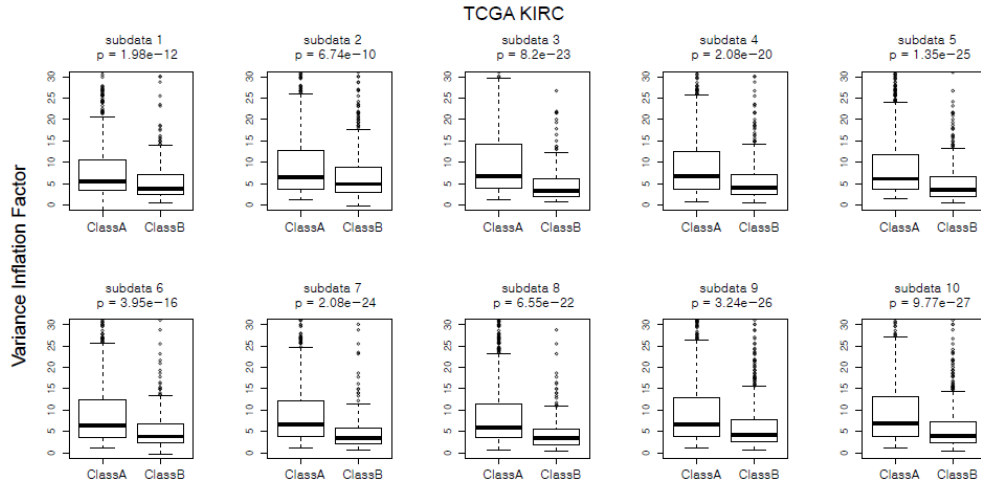


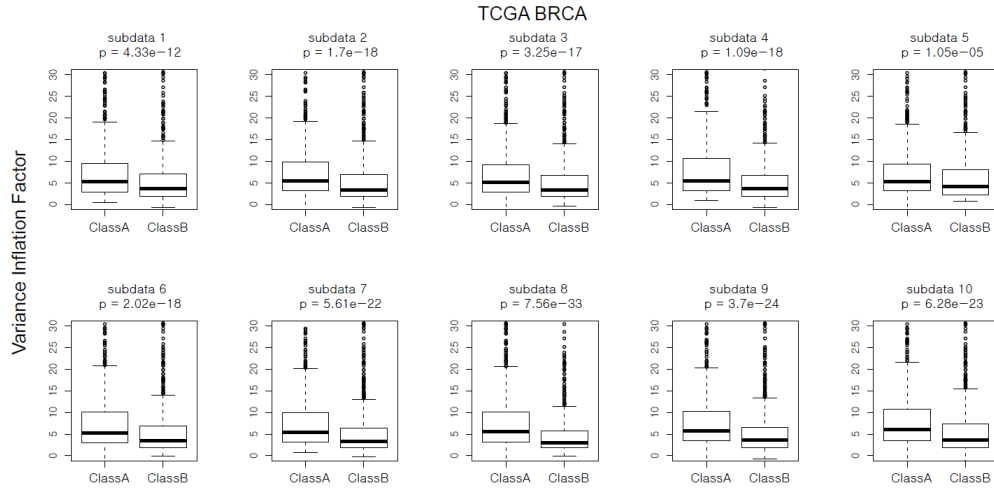
Figure S2.2. Average receiver operating characteristic (ROC) curves for two sample cases.

The average ROC curves of the twelve gene-permuting GSEA methods applied to simulation data where inter-gene correlation was 0.3 and the number of replicates were two.

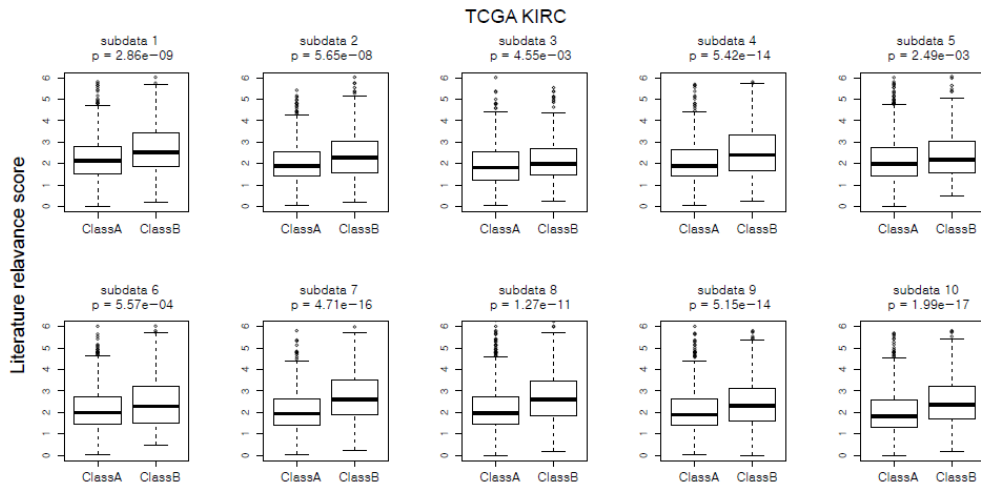
(a)



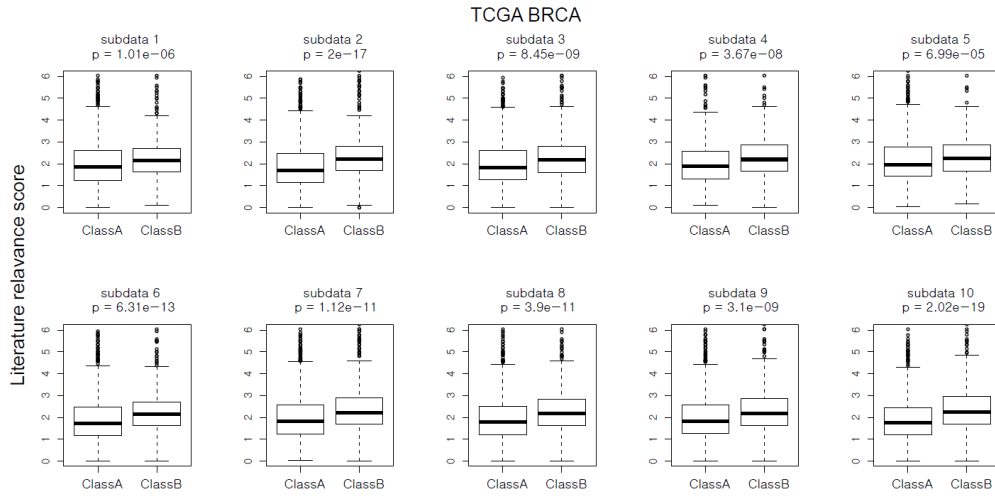
(b)



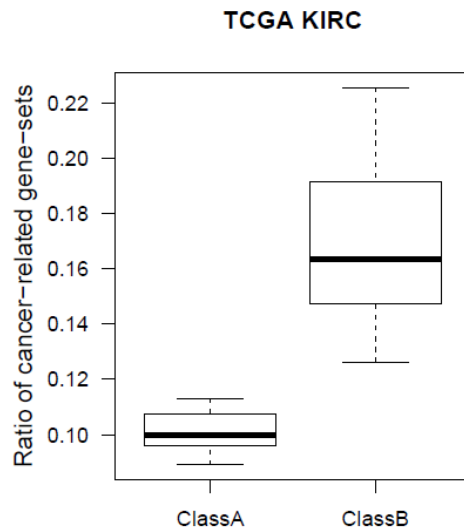
(c)



(d)



(e)



(f)

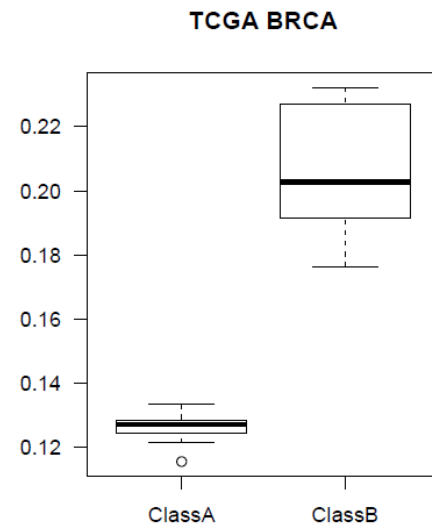


Figure S2.3. The effect of absolute gene-permuting GSEA

Five tumor and matched normal samples were randomly selected from (a,c,e) TCGA KIRC or (b,d,f) BRCA RNA-seq dataset, and original or absolute gene-permuting GSEA were performed (gene score: logFC). (a,b) The distributions of variance inflation factor and (c,d) literature score of gene-sets that were significantly detected ($FDR < 0.25$) only in the original GSEA (ClassA) and both in the original and absolute GSEA (ClassB) were compared (Wilcoxon ranksum test). This process was repeated ten times. (e,f) In addition, the ratio of gene-sets containing terms such as ‘cancer’, ‘tumor’, or ‘carcinoma’ were compared between class A and B.

Two-tailed absolute GSEA produces more false positive results than one-tailed absolute GSEA.

We compared the filtering results by one-tailed and two-tailed absolute GSEA in analyzing Pickrell¹¹² and Li¹²⁰ data. Two-tailed absolute GSEA generated more significant gene-sets than one-tailed absolute GSEA. For example, the GSEA-GP with one-tailed absolute filtering of Pickrell data (gene score: moderated-t) resulted in 2.6 significant gene sets (FDR<0.25) including one true term (chryq11) on average, while that of two-tailed filtering yielded 3.3 significant gene sets including one true term on average. When logFC was used as gene score, the one-tailed and two-tailed absolute filtering produced 3.5 and 3.7 significant terms, respectively, including one true term.

Similar result was observed for the Li data. The GSEA-GP with one-tailed absolute filtering detected 8 significant gene sets (FDR<0.1) with three ‘androgen’-related gene sets as shown in the Table 2.1. However, when the two-tailed absolute filtering was applied, it detected 14 significant gene sets including the same three androgen-related terms. When logFC was used as the gene score, the one-tailed and two-tailed absolute filtering detected 242 and 256 significant terms, respectively, including four androgen-related terms. These results imply that one-tailed absolute GSEA yields a little more conservative results.

Chapter III: A powerful pathway enrichment and network analysis tool for GWAS summary data

3.1 Abstract

Pathway-based analysis methods in genome-wide association study (GWAS) are being widely used to uncover novel multi-genic associations. Many of the pathway-based methods tested the enrichment of the associated genes in the pathways, but exhibited low powers and were highly affected by free parameters. A novel standalone software GSA-SNP2 was developed in this study for pathway enrichment analysis of GWAS p -value data. GSA-SNP2 provides high power, decent type I error control, and fast computation by incorporating the random set model and the SNP-count adjusted gene score. In a comparative study using simulated and three real GWAS data, GSA-SNP2 exhibited high power and best discriminatory ability compared to six existing enrichment-based methods and two *self-contained* methods which is an alternative pathway analysis approach. Based on these results, the differences between pathway analysis approaches, and the effects of different correlation structures on the pathway analysis were also discussed. In addition, GSA-SNP2 visualizes protein interaction networks within and across the significant pathways so that the user can prioritize the core subnetworks for further mechanistic study.

3.2 Introduction

Improving the power of genome-wide association study (GWAS) has been a big challenge for the last decade. After the multiple testing correction, only a handful of SNP markers were obtained in a typical GWAS. Analysis of such top-ranked SNPs discarding all except ‘the tip of the iceberg’ was capable of revealing only a few associated functions. As the sequencing cost keeps dropping, the whole genome sequencing data are being used for GWAS¹²¹ which poses a much greater multiple testing burden. To address the problem, a number of multi-loci (gene or pathway) based association analysis methods were developed. These methods substantially increased statistical power, and revealed many novel genes and pathways that were not found by the single SNP-based approach¹²²⁻¹²⁴. In particular, pathway-based association analysis methods directly provide biologic interpretations, and are capable of detecting aggregate association of multiple genes even when the individual genes are only moderately associated. In earlier times, most of the pathway-based GWAS analysis methods incorporated competitive null hypothesis⁹⁹ and tested the relative enrichment of the associated genes in each pathway gene-set. GenGen¹²⁵, GSEA-SNP¹²⁶, iGSEA4GWAS¹²⁷, SSEA¹²⁸ and MAGENTA⁶² implemented modified

GSEA algorithms which were originally developed for the pathway analysis of gene expression data, GSA-SNP⁶¹ implemented modified Z-test as well as two GSEA algorithms, Aligator¹²⁹ and Gowinda⁶⁴ provided Gene Ontology over-representation analysis accounting for the gene size (or SNP count), INRICH⁶³ tested enrichment of pathway gene-sets across independent genomic intervals, and MAGMA⁶⁵ exploited multiple regression models on gene and gene-set analysis. Whereas competitive methods for GWAS data provided fast and simple implementations, many of them exhibited low powers and were susceptible to some free parameters.

The pathway-based association analysis methods were then developed for *self-contained* null hypothesis in recent years^{68, 99, 123, 130}. Competitive methods directly target pathway-level aberrations by testing the *enrichment* of the associated genes within each pathway, whereas self-contained methods test the *existence* of the associated genes therein¹⁰³. Thus, self-contained methods are in general highly sensitive, so are useful in discovering novel pathways. However, genes typically have multiple functions and mere existence of associated gene(s) does not always imply a *pathway-level* aberration. So, both approaches are useful and complementary to each other.

Besides, protein-protein interaction (PPI) networks were also considered for analyzing GWAS summary data to identify large modules of associated proteins beyond the pre-defined pathway gene-sets¹³¹⁻¹³². In this way, interrogation of GWAS data from different levels of biologic objects (SNP, gene, pathway and network) has proven useful for revealing novel associations to the phenotype of interest. Here, a novel C++ standalone tool GSA-SNP2 is presented that accepts GWAS SNP *p*-values and implements a powerful competitive pathway analysis as well as PPI network visualization in the significant pathways. Compared to its previous version⁶¹, GSA-SNP2 provides much improved type I error control by using the SNP-count adjusted gene scores, while preserving high statistical power. The gene scores are adjusted for the SNP counts for each gene using monotone cubic spline trend curve. It was critical to remove high scoring (potentially associated) genes before the curve fitting to achieve high power. The performance of GSA-SNP2 was compared with those of six existing competitive pathway analysis methods and two recently developed self-contained method using simulated GWAS data and DIAGRAM consortium data (type II diabetes). Based on these results, the difference between pathway analysis approaches for GWAS data, and the difference in gene correlation structures between GWAS and gene expression data and their effects on competitive pathway analysis were also discussed. GSA-SNP2 visualizes the PPI networks within (local) and across (global) the significant pathways. These networks suggest how the key proteins interact to each other and affect their neighbors in the aberrant pathways. The global network, in particular, shows the core PPI structure that cannot be represented by single pathways suggesting clues for mechanistic study. GSA-SNP2 is freely available at <https://sourceforge.net/projects/gsasnp2>.

3.3 Materials and Methods

3.3.1 Algorithm of GSA-SNP2

GSA-SNP2 employs Z-statistic in evaluating gene sets (pathways) like GSA-SNP. The critical improvement is resulted from the usage of the gene scores adjusted for the SNP counts for each gene using monotone cubic spline trend.

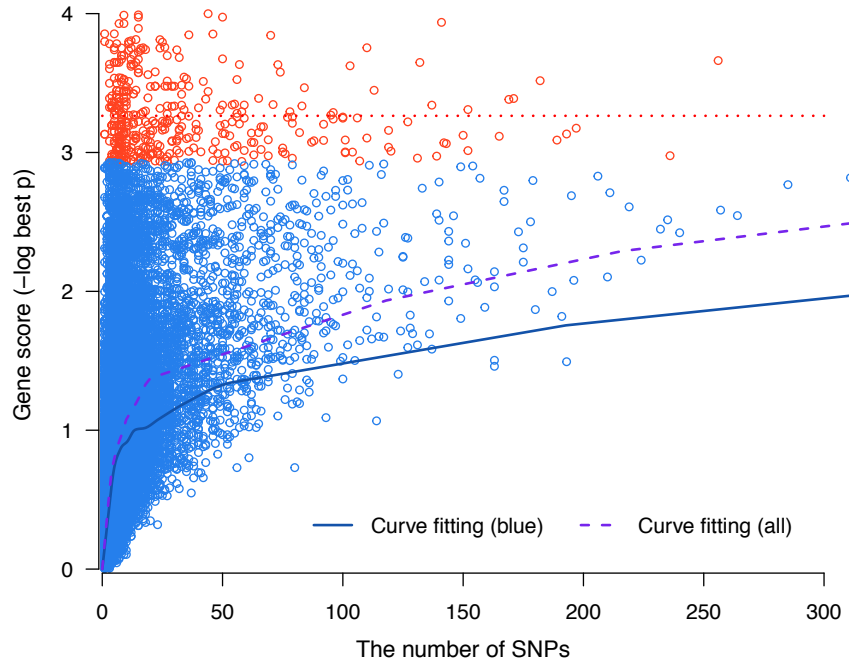


Figure 3.1. The monotone cubic spline trend curves.

Red points represent high scoring genes that have zero correlation coefficient (red dotted line). Both the trend curves with (purple) or without (blue) red points are represented. The blue curve is used for calculating the adjusted gene scores.

Adjusted gene scores

SNPs that are located in the range of a gene [gene start – padding, gene end+padding] are assigned to the gene, where the padding size of a gene is chosen among {0, 10000, 20000}. Then, the initial gene score is given as the maximum of $-\log(\text{SNP } p\text{-value})$ for those SNPs. These gene scores in general tend to increase as the number of assigned SNPs is increased. Thus, the initial gene scores are adjusted for the number of assigned SNPs using monotone cubic spline trend as shown in Figure 3.1. Many genes had very high scores irrespective of the increasing trend for the SNP counts, so such high scoring (presumably associated) genes are removed before fitting the trend curve. In other words, a range of top gene scores is searched so that their correlation coefficient becomes zero (red points) and the

corresponding genes are removed. And then, a monotone cubic spline curve (blue solid curve) is fitted for the remaining gene scores (blue points). Note that without filtering such high scoring genes, the trend curve was rather enhanced (purple dash), and the power of our method was much lowered eventually.

The adjusted gene score for i th gene g_i is given as

$$Adj(g_i) = -\log(p_i) - C(g_i)$$

, where p_i is the best p -value among the SNPs assigned to g_i and $C(g_i)$ is the estimated gene score on the trend curve. Note that the removal of the high scoring genes is only for the curve fitting and they are all restored when calculating the adjusted gene scores. See Supplementary information of Chapter III for the detailed algorithms for outlier treatment, conversion to monotonic data, curve fitting process.

Pathway statistic

Given a list $S = \{S_1, S_2, \dots, S_K\}$ of K gene-sets (pathways), each gene-set S_i ($0 \leq i < K$) can be assessed by Z-statistic as follows:

$$Z(S_i) = \frac{X_i - m}{\sigma / \sqrt{N_i}}$$

, where X_i is the average of the *adjusted* gene scores in the gene-set S_i , m and σ are respectively the mean and the standard deviation of all the adjusted gene scores, N_i is the number of genes in the current gene-set. *Random set* theory¹⁰⁰ is implemented in GSA-SNP2 to capture more closely the impact of the set size on the set score. Under the light of random set model, the mean m does not depend on attributes of the gene-set, but the standard deviation σ^* depends on the gene-set size N_i :

$$\sigma^* = \sigma \cdot \left(\frac{|G| - N_i}{|G| - 1} \right)^{\frac{1}{2}}$$

, where $|G|$ is the total number of genes analyzed. The final gene-set statistic is as follows:

$$Z(S_i) = \frac{X_i - m}{\sigma^* / \sqrt{N_i}}$$

Adjacent gene filtering

Some of the genes in a pathway are closely located on the genome or highly overlapping family genes, and some of them may belong to the same linkage disequilibrium (LD) block. Such genes exhibit a positive correlation in their p -values and may contribute to increasing false positive pathways. To prevent this possibility, the adjacently located genes in a pathway are alternatively removed if they have high positive genotype correlations (>0.5) in the 1000 genome data. See Supplementary Data for the detailed algorithm. However, in practice, only a small portion of genes in a pathway were adjacently

located and at the same time had high correlations. As a result, this filtering process had little effect in reducing false positives in the type I error control in our tests.

3.3.2 Competitive pathway analysis tools

The type 1 error rate control and statistical power of GSA-SNP2 was compared with other existing competitive pathway analysis methods that analyze GWAS summary data (Z-test of GSA-SNP (GSA-SNP1), iGSEA4GWAS, MAGMA, MAGENTA, INRICH and Gowinda). MAGMA was tested for mean (MAGMA-mean) and top1(MAGMA-top1) SNP statistics as well as their combination (MAGMA-multi). For MAGENTA, two default enrichment cutoffs (75 and 95 percentiles of all gene scores) were used. For INRICH, the SNP intervals were constructed for top 1% association p-value. $R^2=0.5$ was used for another LD-clumping parameter. Gowinda was tested for gene-mode and candidate SNPs were selected for top 1%, 5% or 10% association p-value. Other parameters were set as default.

3.3.3 Simulation study

The genotypes of 10000 individuals were simulated by randomly pairing the haplotypes of 1000 Genome European samples. The phenotype Y of each individual was calculated based on the linear model. For type 1 error rate control test, following model was used.

$$Y = \beta_1 X_1 + \cdots \beta_k X_k + \varepsilon$$

where X_1, \dots, X_k are normalized additive genotypes of k effective SNPs, β_1, \dots, β_k are SNP effect (set as one in this study) and ε is residual with $\varepsilon \sim N(0, \sigma^2)$. In the type 1 error rate test, 300 effective SNPs were randomly selected within gene region. The phenotype variance σ^2 is determined by the narrow-sense heritability (h^2). In this case, the simulation data were generated for $h^2=25\%$ or 50% .

For power test, following model was used.

$$Y = \beta_1 X_1 + \cdots \beta_k X_k + \gamma(G_1 + \cdots + G_M) + \varepsilon$$

where γ is gene-set effect and G_1, \dots, G_M are gene effects of M causal genes in the target gene-set. The gene effect of a gene g (G_g) is defined as $G_g = (X_{g_1} + \cdots + X_{g_L})/\sqrt{L}$ where X_{g_1}, \dots, X_{g_L} are normalized additive genotypes of L causal SNPs within gene g . In this case, the total heritability was decomposed into the background heritability ($h_b^2 = \frac{\text{Var}(\beta_1 X_1 + \cdots \beta_k X_k)}{\text{Var}(Y)}$) and gene-set specific heritability ($h_g^2 = \frac{\text{Var}(\gamma(G_1 + \cdots + G_M))}{\text{Var}(Y)}$), assuming that X_1, \dots, X_k and G_1, \dots, G_M have no correlation. Gene-set effect γ and phenotype variance σ^2 is determined by the combination of h_b^2 and h_g^2 . The power simulation data were generated for $h_b^2 = 25\% \text{ or } 50\%$ and $h_g^2 = 4\% \text{ or } 8\%$. In this case, 100 background SNPs were randomly selected within the gene regions, and 10~40% of causal genes in a target pathway were randomly chosen. For each causal gene, one causal SNP was randomly assigned. 674 Reactome pathways (set size: 10~200) were used in the simulation test^{42, 133}.

3.4 Results and Discussion

3.4.1 Type I error rate simulation test

False positive (FP) control test was repeated 20 times for each condition and figure 3.2 shows the number of FP gene-sets ($FDR < 0.05$) detected by each method. Because the causal SNPs were randomly distributed on the genome, none of gene-sets were enriched under this simulation setting. However, GSA-SNP1 and iGSEA4GWAS detected many FP sets (Median FP count of iGSEA4GWAS: 59.5 for $h^2=25\%$, 42 for $h^2=50\%$; GSA-SNP1: 26 for $h^2=25\%$, 35.5 for $h^2=50\%$ out of 674 pathways). Other methods were good at FP control. GSA-SNP2 showed highly improved FP control compared to GSA-SNP1 by applying adjusted gene scoring method. GSA-SNP2 was slightly liberal than INRICH, MAGMA and MAGENTA that detected almost no FP sets, but it still showed quite decent false positive control. Its median count of FP set was merely 2 for $h^2=25\%$ and 1 for $h^2=50\%$. The results from Gowinda varied according to the SNP p-value cutoff. The false positive rate increased as the p-value cutoff increased. It showed best FP control with top 1% SNPs, but some data generated high FP sets under this condition (39/674, 5.8%).

3.4.2 Power simulation test

The statistical power of each method was tested for the combination of two background heritability ($h_b^2=25\%$, 50%) and two set-specific heritability ($h_g^2=4\%$, 8%). One target pathway was assigned for each simulation and I tested how many target pathways were significantly detected ($FDR < 0.05$) among 50 trials. GSA-SNP1 and iGSEA4GWAS were excluded from the test because they were vulnerable to the false positive control. Figure 3.3 shows the power of each method for each condition. GSA-SNP2 showed the best power for all conditions ($h_b^2 = 25\%/h_g^2 = 8\%$: 78.0%, $h_b^2 = 25\%/h_g^2 = 4\%$: 60.0%, $h_b^2 = 50\%/h_g^2 = 8\%$: 65.3%, $h_b^2 = 50\%/h_g^2 = 4\%$: 44.0%). The power of MAGMA varied according to the analysis model. In most cases, MAGMA-top1 showed slightly better power than MAGMA-mean. Their combination (MAGMA-multi) considerably improved the true positive detection compared to either method, but still its power was quite lower than that of GSA-SNP2 (best power: 54.5% at $h_b^2 = 25\%/h_g^2 = 8\%$). INRICH and MAGENTA exhibited low powers compared to other methods (best power of INRICH: 16.3%, MAGENTA (75%): 22.0%, MAGENTA (95%): 14.3% at $h_b^2 = 25\%/h_g^2 = 4\%$). The results from Gowinda varied according to the SNP p-value cutoffs. Among three cases, Gowinda showed best power using top 1% SNPs as the candidate SNPs (best power: 33.3% at $h_b^2 = 25\%/h_g^2 = 8\%$).

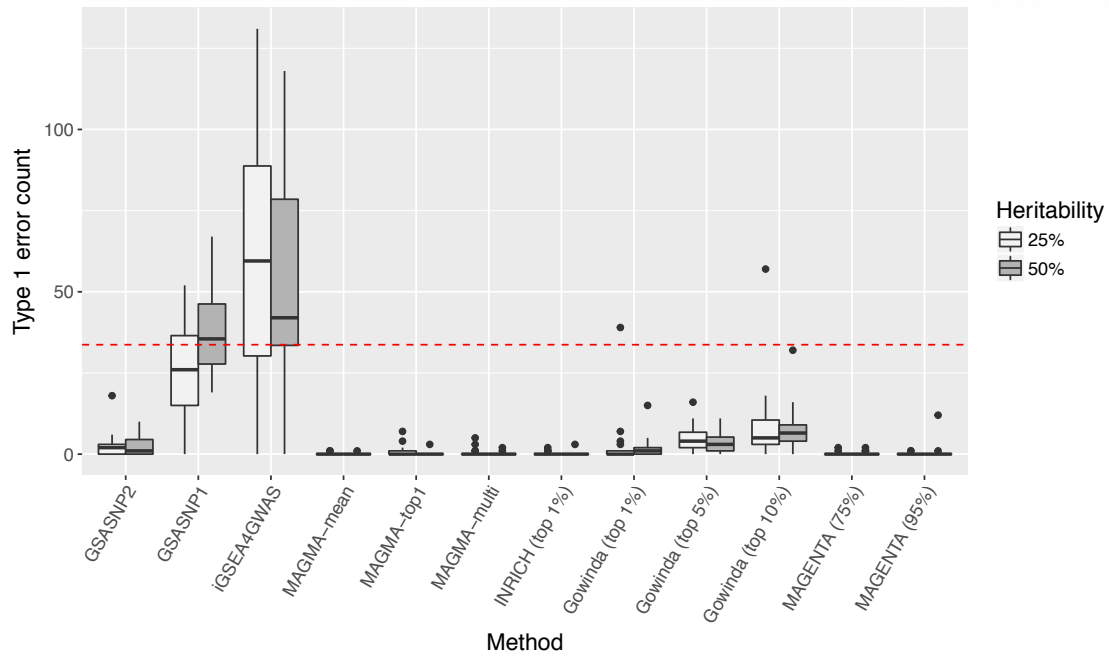


Figure 3.2. Type 1 error rate comparison.

The boxplots of the false positive gene-sets count ($FDR < 0.05$) detected by six competitive pathway analysis methods are shown. The simulation was performed for two heritability values (25% and 50%) and each simulation was repeated 20 times. MAGMA was tested for three gene models, INRICH was tested for approximate top 1% of SNPs, Gowinda was tested for approximate top 1%, 5% and 10% and MAGENTA was tested for 75% and 95% of enrichment cutoff. The red line indicates the 5% of total pathways.

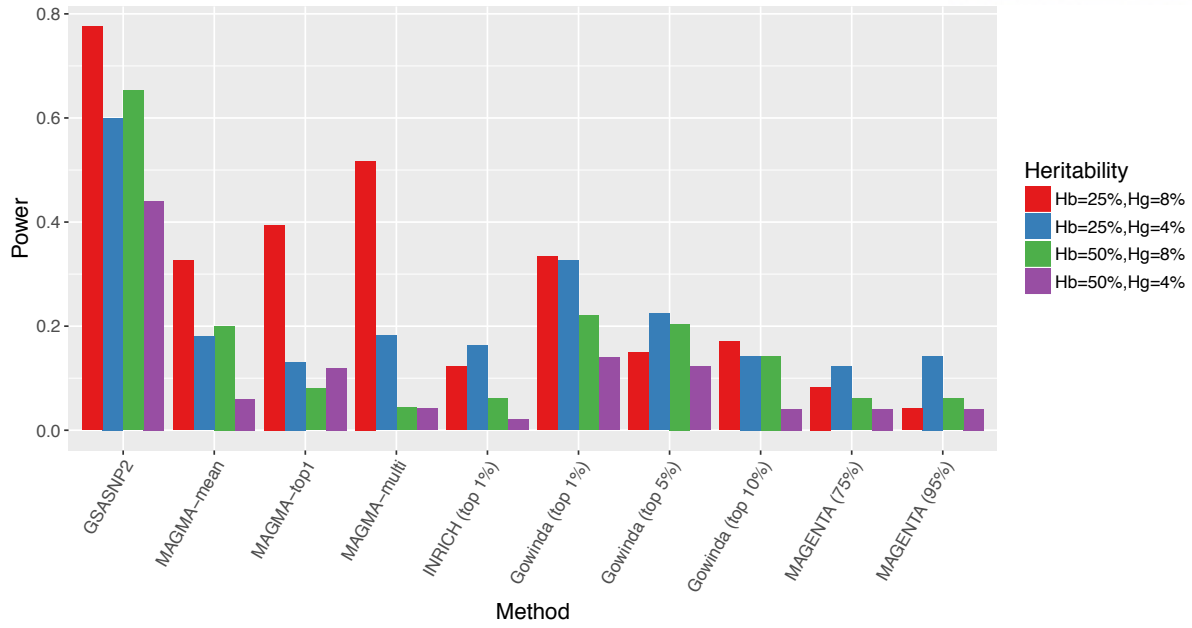


Figure 3.3. Statistical power comparison.

The power of each competitive pathway analysis method under the four simulation settings ($h_b^2 = 25\%/h_g^2 = 8\%$, $h_b^2 = 25\%/h_g^2 = 4\%$, $h_b^2 = 50\%/h_g^2 = 8\%$ and $h_b^2 = 50\%/h_g^2 = 4\%$) are represented. The parameters of each method were same with those used in the type 1 error rate test.

3.4.3 Performance comparison using real data

The performance of GSA-SNP2 was compared with multiple competitive (GSA-SNP, INRICH ($p=1E-6$, and $p=1E-8$), Gowinda ($p=1E-3$, $p=1E-2$ and $p=5E-2$), iGSEA4GWAS, MAGENTA (enrichment cutoff: 75% and 95% gene score), MAGMA-mean, MAGMA-top1 and MAGMA-multi) and self-contained (sARTP and self-contained MAGMA) methods using three public data. I also included GSA-SNP2 applied with VEGAS2 gene scores in the comparison (GS2VEGAS-all and GS2VEGAS-top1: all or best SNP(s) in a gene region was(were) used for gene score evaluation).

First, the DIAGRAM consortium stage 1 GWAS p-values were used for comparing the statistical power. 16 curated type II diabetes (T2D) related pathways¹³⁴ as well as those including the word ‘diabetes’ were regarded as gold standard gold standards and were summarized into 15 categories (Table 3.1). All the mSigDB C2 pathway gene-sets that correspond to these categories were listed in Supplementary Information of this chapter (denoted TP pathways). Figure 3.4 shows the comparison results between different methods: the cumulative gold standard pathway count for each pathway rank were plotted for each method up to q-value<0.25. Same graph with strict cut-off (q-value<0.05) is shown in figure S3.2. See also Table S1 of its original paper published in NAR⁶⁷ for the detailed results for each method compared. GSA-SNP2 exhibited high power and outperformed the other competitive and MAGMA

self-contained methods in the overall TP rank distribution. It was also showed slightly better gold standard rank distribution compared with powerful self-contained method sARTP. Except GSA-SNP2, GSA-SNP and iGSEA4GWAS, other competitive methods detected only small number of gold standard pathways (≤ 15) due to the low power. GSA-SNP2, GSA-SNP and iGSEA4GWAS detected 41, 47 and 49 TP pathways among 108, 232 and 240 significant pathways ($FDR < 0.25$), respectively. All self-contained methods exhibited high power. sARTP detected 52 gold standard pathways out of 193 significant terms, and self-contained MAGMA-mean, showing the best precision among MAGMA series, detected 85 gold standard pathways out of 552 significant sets. The gold standard pathways significantly detected by each method were counted for 15 categories (Table 3.1). Because the pathways ranked lower than top 100 may not draw much attention, I counted them up to 100th rank. Here, I focused on four methods which detected more than 25 gold standard pathways (GSA-SNP2, iGSEA4GWAS, sARTP and self-contained MAGMA-multi). Those methods detected all gold standard pathways in ‘regulation of beta cell’ category. Except that, GSA-SNP2 best predicted at ‘diabetes’, ‘blood glucose regulation’, ‘branched chain amino acid metabolism’, ‘inflammation’ and ‘Notch signaling’ pathways; iGSEA4GWAS best predicted at ‘cell cycle’, ‘unfolded protein response’ and ‘glycolysis and gluconeogenesis’ pathways; self-contained MAGMA-multi best predicted at ‘diabetes’, ‘unfolded protein response’, ‘Notch signaling’ and ‘mitochondrial dysfunction’ pathways; and sARTP best predicted as many as six categories such as ‘diabetes’, ‘adipocytokine signaling’, ‘unfolded protein response’, ‘fatty acid metabolism’, ‘PPARG signaling’ and ‘WNT signaling’ pathways. Overall, GSA-SNP2 detected large number of TP terms within top 100 pathways, and showed the comparable coverage of diverse gold standard categories compared with two powerful self-contained methods.

Next, the height GWAS p-values from GIANT consortium 2010 were analyzed¹³⁵. The 15 gold standard pathways related height and bone regulation were curated from three independent studies. First, Pers et al. performed DEPICT pathway analysis using large size of height GWAS data from GIANT consortium 2012-2015 (sample size: 253,288)¹³⁶⁻¹³⁷. Because large sample size increases the statistical power, and DEPICT properly controls the type 1 error-rate, it was regarded as a good source for examining the height-related pathways. From 183 significant pathways ($FDR < 0.01$), 12 gold standard categories were found reported in the publications such as skeletal system development and epigenetics¹³⁸⁻¹³⁹. Second, Marouii et al. analyzed rare and low-frequency coding variant that affected to human adult height, and suggested several height-associated genes and pathways¹⁴⁰. Among them, ‘proteoglycan’ and ‘reactive oxygen species’ were experimentally validated in other studies, so those were included in the gold standard categories¹⁴¹⁻¹⁴². Third, ‘telomerase activity’ that have important role in chondrocyte proliferation during bone elongation was also included in the gold standard categories¹⁴³. The 15 height-related gold standard categories and related mSigDB C5 gene ontology terms (v 6.0) are listed in the Supplementary information of this chapter. The detailed analysis result of

all methods except sARTP are represented in the Table S2 of its original paper⁶⁷. sARTP was not tested with height data, because it cannot be applied to quantitative trait GWAS data. In this case, the cumulative gold standard pathway counts were plotted up to $q\text{-value} < 0.05$ because most competitive methods showed greatly increased power compared to previous example due to the large sample size (183,727; figure 3.4). Similar to the previous case, GSA-SNP2 exhibited the high power and the best gold standard pathway prioritization. It detected 50 TP pathways out of top 100 significant terms. Other GSA-SNP series methods including GSA-SNP and GSA-SNP2 applied with VEGAS2 gene scores (GS2VEGAS-mean, GS2VEGAS-top1) also showed outstanding power and TP pathway prioritization compared to other methods. GSA-SNP, GS2VEGAS-mean and GS2VEGAS-top1 detected 46, 50 and 53 TP pathways out of top 100 pathways, respectively. Unlike previous example, where relatively small number of samples (69,033) were used, many competitive methods including MAGMA, MAGENTA (95%) and Gowinda showed highly increased power in this case. Especially, MAGENTA and MAGMA exhibited better TP prioritization compared to self-contained MAGMA methods (MAGENTA detected 35 TP pathways out of 73 significant terms; MAGMA-multi detected 40 TP pathways out of top 100 pathways; and self-contained MAGMA-multi detected 37 gold standard pathways out of top 100 pathways). There was difference in preferred TP categories for each method. For example, the competitive MAGMA methods detected the largest number of ‘skeletal system development’ pathways such as cartilage and chondrocyte development (e.g., MAGMA-multi detected 23 related terms), and many of them were in the top ranking. They were also top-ranked in the MAGENTA result. On the other hand, GSA-SNP series detected the majority of ‘epigenetics’ pathways (21~22 related terms were detected by all GSA-SNP series), and many of them were placed in the top ranking. All GSA-SNP series also specifically detected many ‘telomerase activity’ pathways within top 100 terms. The most ‘insulin-like growth factor and growth hormone’ pathways were detected by GS2VEGAS (six terms were detected by GS2VEGAS-all while other methods detected three or less terms).

I also compared the statistical power using Korean height GWAS p-values where relatively small samples (8,842) were used¹⁴⁴. In this case, the cumulative gold standard pathway counts were plotted up to $q\text{-value} < 0.25$ due to the lowered powers in overall methods. See figure S3.2 where same graph was drawn for $q\text{-value} < 0.05$. Again, GSA-SNP2 showed high power and outstanding gold standard pathway prioritization compared to other methods. It detected 44 gold standard pathways out of top 100 terms. Here, GS2VEGAS and MAGENTA (75%) methods showed slightly better gold standard rank than GSA-SNP2 in the front (~40th rank). Although MAGENTA had low power, it exhibited the highest gold standard pathway density (25 out of 41 significant terms; 61.0%) showing its great false positive control. The statistical powers of MAGMA methods were severely decreased compared to GIANT height case. MAGMA-mean and MAGMA-multi detected no significant pathways and only MAGMA-top1 detected five ‘skeletal system development’ and one ‘epigenetics’ pathways. It implies that

MAGMA is quite sensitive to GWAS sample size than other methods. The preferred gold standard categories were similar to the GIANT height case. For example, self-contained MAGMA methods detected many ‘skeletal system development’ pathways than others (13~16 pathways; GSA-SNP2 detected 11 relevant pathways and others detected eight or less), while GSA-SNP series detected particularly many ‘epigenetics’ and ‘telomerase’ pathways than others.

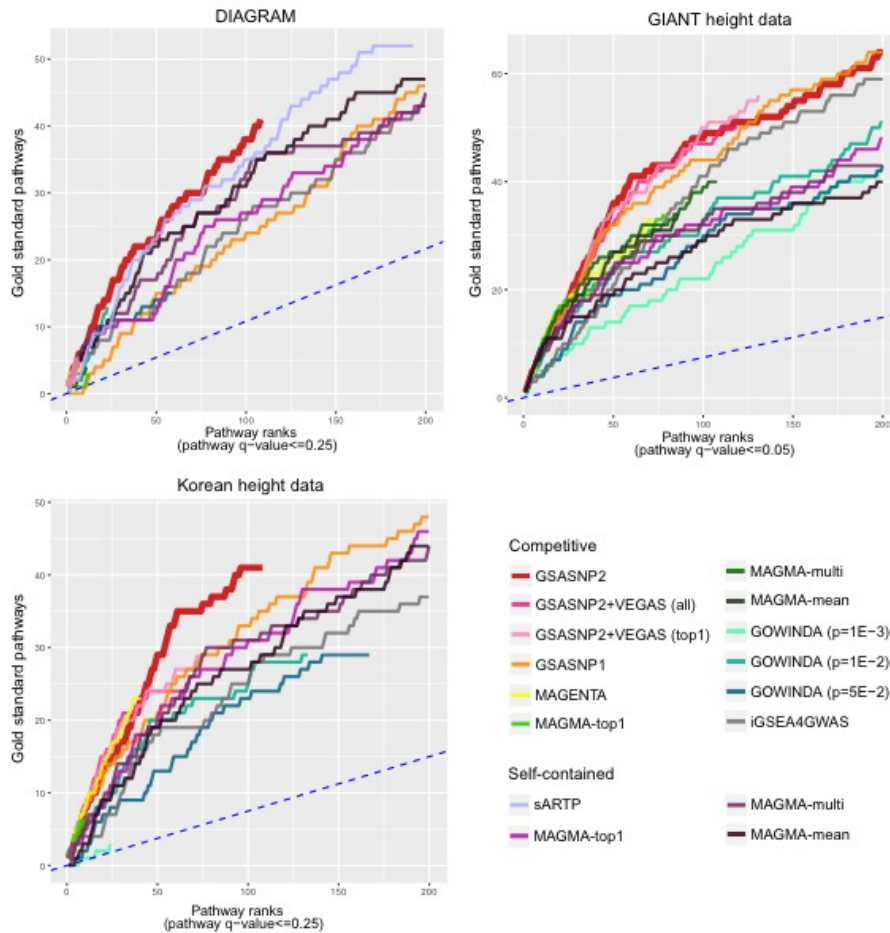


Figure 3.4. Power comparison using real data.

For three public GWAS summary statistics data (DIAGRAM, GIANT height and Korean height data), the cumulative gold standard pathway count distributions of six competitive and two self-contained pathway analysis methods were plotted. The results from INRICH were not represented because it failed to detect more than one gold standard pathway for all cases. The blue dashed line indicates the expected distribution of the cumulative gold standard pathway count.

Table 3.1. Power comparison using canonical pathways for diabetes.

Canonical pathways for diabetes were classified into 15 categories, and the numbers of top 100 canonical pathways were counted for each category. The insignificant pathways (FDR>0.25) were not counted. The total TP counts are indicated in the bottom.

Category	GSASNP2	GS2VEGAS (best P)	GSASNP1	iGSEA4 GWAS	MAGMA- multi	MAGENTA (75%)	GOWINDA (p=0.01)	sARTP (self-contained)	MAGMA-multi (self-contained)
Diabetes	3/4	1/4	3/4	2/4	0/4	1/4	2/4	3/4	3/4
Regulation of beta cell	3/3	3/3	1/3	3/3	0/3	2/3	3/3	3/3	3/3
Insulin/blood glucose level	10/25	0/25	8/25	2/25	0/25	0/25	1/25	4/25	6/25
Adipocytokine signaling	1/6	0/6	0/6	0/6	0/6	0/6	0/6	2/6	1/6
Cell cycle	4/22	1/22	1/22	5/22	2/22	0/22	1/22	4/22	4/22
Circadian rhythm	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6
Unfolded protein response	0/2	0/2	0/2	2/2	2/2	0/2	0/2	2/2	2/2
Branched-chain amino acid metabolism	1/2	0/2	1/2	0/2	0/2	0/2	1/2	0/2	0/2
Fatty acid metabolism	2/10	0/10	3/10	1/10	0/10	1/10	0/10	5/10	3/10
Glycolysis and Gluconeogenesis	0/3	0/3	0/3	2/3	0/3	0/3	1/3	0/3	1/3
Inflammation	8/22	0/22	2/22	3/22	0/22	0/22	0/22	4/22	2/22
NOTCH signaling	6/14	0/14	3/14	5/14	0/14	0/14	4/14	5/14	6/14
PPARG signaling	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1
WNT signaling	0/11	0/11	2/11	1/11	0/11	0/11	0/11	2/11	1/11
Mitochondrial dysfunction	0/5	0/5	0/5	0/5	0/5	0/5	0/5	0/5	1/5
Total TP pathways	38/100	5/7	24/100	26/100	4/11	4/9	13/23	35/100	33/100

3.4.4 Comparison of competitive and self-contained pathway analysis results

GSA-SNP2 and sARTP results were further compared by the pathways exclusively detected by either method. The top ten pathways that were significant with GSA-SNP2 but were least significant with sARTP, and vice versa were selected and the distributions of gene p -values (VEGAS best p option) were compared in Figure 3.5. In the former case, several genes had similar low p -values which seemed to collectively represent the pathway-level aberrations. On the other hand, in the latter case, most pathways contained one or two extreme gene p -values which seemed to dominate those pathways. If such extreme genes also belong to many other pathways, the association of the corresponding pathway may not be very reliable. Although competitive methods are also affected by such outlier genes, and self-contained methods are also capable of detecting pathways composed of moderately associated genes only, these examples demonstrate the difference of the two GWAS pathway analysis approaches.

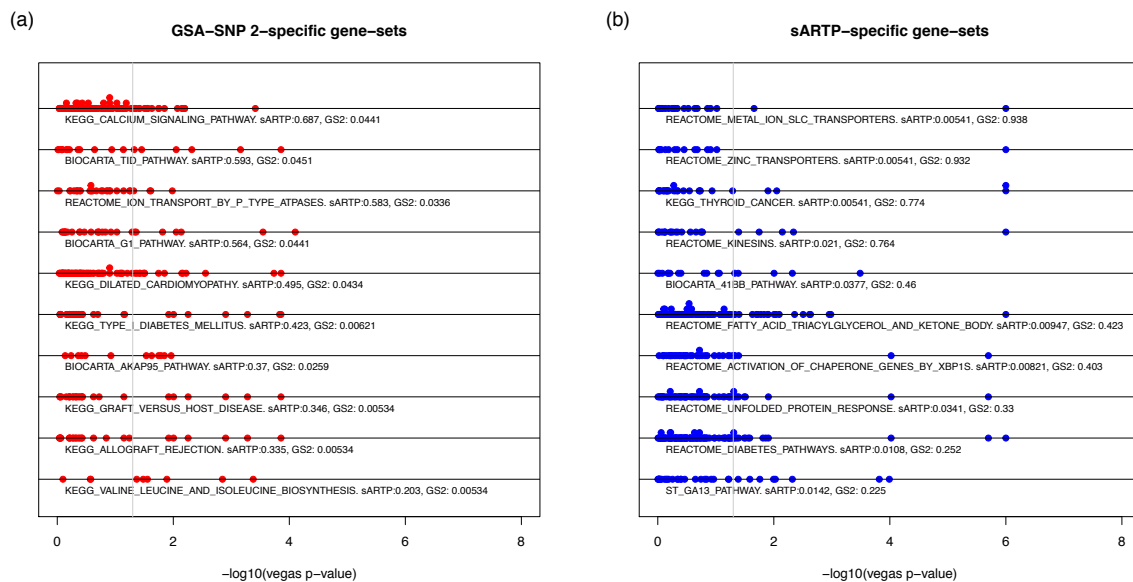


Figure 3.5. Comparison of gene p -value distributions in the pathways that are only significant with (a) GSA-SNP2 or (b) sARTP.

3.4.5 Comparison with the competitive pathway analysis for gene expression data

The core algorithms used for competitive pathway analysis of GWAS data are virtually the same as those used for gene expression data. It is well known that the competitive methods for gene expression data suffer from inflated type I errors caused by the inter-gene correlations in each pathway^{99, 101}. Interestingly, in the test for GWAS summary data, competitive methods mostly resulted in little false positives. There is a substantial difference in the inter-gene correlation structure in each pathway

3.4.6. Comparison of running times

Lastly, the running times for each software were compared for the DIAGRAM data and the C2 canonical pathway sets (Table 3.2). GSA-SNP, GSA-SNP2, MAGMA-mean, INRICH and Gowinda were quite fast taking only a few minutes, while sARTP took over ten days run on the same PC (Intel Xeon Processor X5670 @ 2.93GHz, 12 CPUs and 24GB of RAM).

Table 3.2. Running times for seven pathway analysis programs for GWAS summary data.

Method	Time	Permutation
GSA-SNP2 (command ver.)	1.53 min	-
GSA-SNP	1.49 min.	-
MAGMA-mean	3.03 min	-
MAGMA-top1	34.85 min	-
MAGMA-multi	41.85 min	-
i-GSEA4GWAS	30 min.	-
MAGENTA	114.18 min	10000
Gowinda (p=0.001)	0.62 min.	10000
Gowinda (p=0.01)	0.80 min.	10000
Gowinda (p=0.05)	2.01 min.	10000
INRICH (p1=1E-6)	0.85 min.	10000
INRICH (p1=1E-4)	2.41 min.	10000
sARTP	10.41 days	100000

3.4.6 Network visualization

GSA-SNP2 visualizes protein interaction networks within individual and across significant pathways. Network plots are generated based on STRING¹⁴⁵ or HIPPIE⁷⁵ networks, and the cut-offs for gene and pathway scores for visualization are selected by the user. Clicking on the gene node pops up a table which shows the gene name, mapped SNPs, the neighboring genes, their association scores as well as further detailed information via the hyperlink to outer databases such as GeneCards¹⁴⁶ and dbSNP¹⁴⁷. The network data are also provided as a text file which also shows the pathways that contain the interacting protein pairs.

In particular, the global network can provide interacting protein pairs that do *not* belong to any of the single pathways. Such protein pairs may have drawn relatively less attention, but can provide useful information for mechanistic study. For example, the global network (extracted from HIPPIE network)

of the significant pathways (FDR<25%, gene score<0.01) obtained from DIAGRAM data contained a sub-network composed of eight genes such as TNF, RAB5A, CHUK, LTA, CARS, IGF2BP2, HSPA1L and HSPA1A (Figure 3.6). Among them, TNF and RAB5A have been individually studied and both are known to regulate the insulin-responsive glucose transporter (GLUT4)¹⁴⁸⁻¹⁵⁰, a key protein that regulates the concentration of blood glucose by transporting it to muscle or fat cell. Thus, the deregulation of GLUT4 can lead to insulin-resistance and T2D¹⁵¹⁻¹⁵². The global network shows the two proteins have a medium level of interaction score 0.63 (affinity chromatography technology), and their interaction may have an important implication in T2D.

The DIAGRAM data were also analyzed using STRING network. It provided much denser interaction networks among the high scoring proteins than those for HIPPIE network, and the key T2D proteins TNF and PPARG were represented as hub proteins. Note that many of the interaction edges from STRING network were generated from the literature only which included GWAS papers, and should be carefully analyzed to avoid circular argument.

3.5 Conclusion

GSA-SNP2 is a powerful and efficient tool for pathway enrichment analysis of GWAS summary data. It provides both local and global protein interaction networks in the associated pathways, and may facilitate integrated pathway and network analysis of GWAS data. The five features of GSA-SNP2 are summarized as follows:

- 1) Reasonable type I error control by incorporating gene scores adjusted to the corresponding SNP counts using monotone cubic spline trend.
- 2) High power and fast computation based on the random set model.
- 3) Without any critical free parameter
- 4) Protein interaction networks among the member genes were visualized for the significant pathways. This function enables the user to prioritize core sub-networks within and across significant pathways. STRING and HIPPIE networks are currently provided.
- 5) Easy to use: Only requires GWAS summary data (or gene *p*-values) and takes **only a minute or two** to get results. Other powerful self-contained pathway tools also require SNP correlation input and take much longer time.

3.6 Supplementary information of Chapter III

In this supplementary information, the ‘Adjusting algorithm’ part was written by Dr. Hai C T Nguyen, a co-first author of this research. I generated figures and wrote ‘15 biological processes related to height regulation’ part.

Adjusting algorithm

GSA-SNP2 still employs the Z-statistic method in evaluating gene sets/pathways like GSA-SNP. The critical improvement is the usage of the *adjusted* gene scores instead of the direct $-\log(p\text{-values})$.

Z-statistic method

Assume that there are a list S of K gene sets/pathways, each gene set S_i ($0 \leq i < K$) is assessed by a set score s_i following the Z-statistic:

$$s_i = z(S_i) = \frac{X_i - m}{\sigma / \sqrt{N_i}}$$

, where X_i is the average of the gene scores in gene set S_i , m and σ are respectively the mean and the standard deviation of all the gene scores, N_i is the number of genes in the current gene set. In this work, GSA-SNP2 considers the best p-value among all SNP p-values p_k ($0 \leq k < |G_j|$) in a gene G_j as the assessed gene score g_j ($0 \leq j < |S_i|$). *Random set* theory is also implemented to capture more closely the impact of the set size on the set score¹⁰⁰. Under the light of random set model, the mean m does not depend on attributes of the category (gene set/pathway), though the variance σ^* depends on the category size N_i :

$$\sigma^* = \sigma \cdot \left(\frac{|G| - N_i}{|G| - 1} \right)^{\frac{1}{2}}$$

With the enhancement of random set method, the final set score s_i is modified by the following equation:

$$s_i = z(S_i) = \frac{X_i - m}{\sigma^* / \sqrt{N_i}}$$

However, *adjusted* gene scores are essentially required before GSA-SNP2 can evaluate the set scores.

Zero correlation area detection

Presenting the distribution of gene score over the numbers of SNPs in a gene, zero correlation area is defined by all scores g_i greater than a calculated threshold g_t so that there is not a linear relationship between two mentioned variables. The threshold g_t can be identified by continuously examining

whether the first derivative a of the regressed linear function $f(g_i \geq g_t) = ag_i + b$ is ‘zero’ with various candidates g_t . Because gene scores are discretely distributed over the numbers of SNPs, a ‘zero’ status is flexibly accepted when a is very close to 0. GSA-SNP2 will accept a ‘zero’ status if a is in range $[-1^{-10}, 1^{-10}]$. A ‘zero’ status can also be accepted if there is no better a found in 1,000 continuous trials. To efficiently identify threshold g_t , the fastest searching technique ‘binary search’ is applied. With each $g_t' = \frac{1}{2} (g_{upper} + g_{lower})$, a regressed linear function $f'(g_i \geq g_t) = a'g_i + b'$ is inferred. When conditions of a ‘zero’ status are still not satisfied, the current first derivative a' is checked to decide the setting for the next trial. If $a' < 0$, more data points are required to rise the slope. To add more data points to current candidate area G_{zero} , the new g_t' needs to be degraded. In ‘binary search’ manner where a g_t' is controlled by the upper and lower boundaries, g_{upper} should be degraded first to make the same effect on g_t' . The result is the new g_{upper} is degraded to the level of the old g_t' . Similarly, if $a' > 0$, less data points should be considered in G_{zero} . Sequentially, the g_{lower} needs to be upgraded to a new level. And in ‘binary search’ manner, g_{upper} is assigned to the level of the old g_t' . The new threshold g_t' will obtain the average value of the new setting of g_{upper} and g_{lower} . This procedure is continuously repeated until a ‘zero’ status is reached. In short, the zero correlation area can be detected by the following iteration algorithm:

- Step 0: Begin with a list G of M gene scores: $G = \{g_0, g_1, \dots, g_{M-1}\}^{70}$, which contains all unduplicated gene scores from all gene sets. GSA-SNP2 only considers the list G of valid genes, which are included in at least a gene set/pathway. GSA-SNP2 also only considers the list S of all appropriate gene sets, which contains at least 10 genes and at most 200 genes in default settings. However, GSA-SNP2 provides controllable parameters for gene set size to fit user interests.
- Step 1: The temporary zero correlation threshold g_t' is set to the average of the maximum and the minimum gene scores. $g_t' = \frac{1}{2} (g_{upper} + g_{lower})$. At this point, $g_{upper} = g_{max}$ and $g_{lower} = g_{min}$.
- Step 2: A new candidate for zero correlation area G'_{zero} can be determined by:

$$\{g_i \geq g_t': g_i \in G\}$$

- Step 3: Perform linear regression on current G'_{zero} to obtain the linear function $f'(g_i) = a'g_i + b'$.
- Step 4: Examine whether $f'(g_i)$ satisfies the conditions of a ‘zero’ status. If the conditions are satisfied, the desired zero correlation threshold g_t is found: $g_t = g_t'$; or else, further investigation is required. The first derivative a' is checked to adjust the parameters for the next iteration base on the following criteria:

$$\begin{cases} \text{if } a' < 0, g_{upper} = g_t' \\ \text{if } a' > 0, g_{lower} = g_t' \end{cases}$$

, where g_{upper} and g_{lower} are previously initialized with the maximum and minimum values among available gene scores. In each iteration, only one parameter g_{upper} or g_{lower} is adjusted at a time, the other remains unchanged.

- Step 5: Repeat Step 1-4 until a zero correlation area G_{zero} is discovered, or the best solution is reached. The best solution may not completely satisfy the conditions of a ‘zero’ status but it remains the same continuously for at least 1,000 iterations.

Outlier treatment

In practice, it is found that genes including extremely significant SNP(s) may cause unexpected situation where zero correlation area contains very few or even one data point. In these cases, the extreme gene score is so significant that the regressed first derivative a' is always much greater than 0 over iterations. That a ‘zero’ status is never reached until the extreme data point is left alone makes the detecting zero correlation phase meaningless in most cases. This situation also makes the search progress work in the worst searching case, which takes the longest time. The proposed solution for this problem is treating the extreme values as outliers and taking them out before detecting zero correlation area. Statistical mean m' and standard variance σ' of gene scores corresponding to each number of SNPs are used to identify outliers in each group. A gene score g' is considered as an outlier when $g' \geq m' + 3\sigma'$.

Dual cubic spline estimation

With assumption that gene score tends to increase when the number of SNPs get greater, GSA-SNP2 uses the monotone cubic spline interpolation to fit the binary data after excluding zero correlation area. The flexibility of spline model allows the fitting curve to capture almost any potential non-linear trends. And in theory, the monotonicity characteristic ensures the direction of the trend curve. However, to practically guarantee the monotonicity, it is found that data to be estimated should not be too much fluctuated (Figure 3.1). In other words, the monotonicity should have been implied in the input data. One popular solution is manually select input data points in a monotone manner¹¹⁹: a heuristically fixed number of knots is selected monotonically; and these knots are fed as input into a cubic spline interpolation algorithm. Firstly, the direct drawback of this method is the neglect of an automatic processing framework. The number of knots is decided merely based on expertise experience, which may vary from user to user. Secondly, the missing of automatic also leads to the neglect of adaptation of this method for a wide range of data. To ease the situation, GSA-SNP2 suggests *dual cubic spline estimation* method to automatically and adaptively estimate scattered binary data. The general idea of ‘dual cubic spline estimation’ is to avoid the effect of severe fluctuating data (Figure S3.1) on the fitting curve. Instead of using all data at the same time, GSA-SNP2 classifies data into two strictly monotone groups: monotonically increasing from the minimum and monotonically decreasing from the maximum. By this simple procedure, it is easy to realize that a fitting curve in each group will not be affected by the fluctuation coming from the other group. Further, each group already obeys the monotonicity characteristic itself. At this point, each of the two fitting curves is able to be used as an adjusting trend line to adjust all gene scores in G . However, an extreme selection procedure like that clearly ignores

the integrity and consistency of data. In order to preserve these important characteristics of input data, another simple integration procedure is implemented to ensure each data item having its part contributed. At each knot, the average of inferred data from both curves is used as new data for the final cubic spline estimation. The algorithm can be summarized as followings:

- Step 0: Binary data $D = \{d_i\}$, where $d_i = (x_i, y_i)$, is sorted along x -axis. Generally, it can be assumed that $x_i < x_j$ when $i < j$. And y_{min}, y_{max} are respectively the minimum and the maximum of all y_i . As a consequence, x_{ymin}, x_{ymax} are respectively the x -axis coordination of the minimum and the maximum.
- Step 1: Classify D into two strictly monotone groups D_{upper} and D_{lower} : monotonically increasing from the minimum and monotonically decreasing from the maximum. The first member of D_{upper} is (x_{ymin}, y_{min}) . Assume that the second selected member of D_{upper} is (x_l, y_l) , for any next selected candidate (x_t, y_t) , it is essential to ensure that $y_l > y_{min}$ and $y_t > y_{l-1}$. Similarly, the first member of D_{lower} is (x_{ymax}, y_{max}) . Assume that the second selected member of D_{lower} is (x_n, y_n) , for any next selected candidate (x_{n+1}, y_{n+1}) , it is essential to ensure that $y_n < y_{max}$ and $y_{n+1} < y_n$.
- Step 2: Perform cubic spline interpolation on both D_{upper} and D_{lower} to obtain two monotone curves: C_{upper} and C_{lower} .
- Step 3: For each knot (x_i, y_i) , $0 \leq i < |D|$, inferring the new knot $(x_{i_average}, y_{i_average})$ for the final cubic spline estimation:

$$\begin{cases} x_{i_average} = \frac{1}{2}(x_{i_C_upper} + x_{i_C_lower}) \\ y_{i_average} = \frac{1}{2}(y_{i_C_upper} + y_{i_C_lower}) \end{cases}$$

, where $(x_{i_C_upper}, y_{i_C_upper})$ and $(x_{i_C_lower}, y_{i_C_lower})$ are respectively the inferred data from C_{upper} and C_{lower} .

- Step 4: Perform cubic spline interpolation again on the new integrated data to obtain the final curve C as the adjusting trend line to adjust all gene scores of G . As a result, for each g_i of G , the corresponding *adjusted score* $g_{i_adjusted}$ can be defined as:

$$g_{i_adjusted} = g_i - C(g_i) = g_i - g_{i_C}$$

, where $g_{i_C} = C(g_i)$ is the inferred trend score from the estimated curve C .

Data sampling strategy

Usually, data are sampled at a regular interval of data population (percentile) or data distribution (range). However, when the density of data at a certain area is too high, the samples from the first approach tend to assemble mostly in that area. That fact will limit the general view of the whole data when an estimation is employed. Meanwhile, the second approach treats all areas, whose density are too high or too low, equally. In that way, the meaningful content of data will be simplified while the less important information may be amplified. Unfortunately, distribution of data of gene scores over the numbers of

SNPs is in this bad shape where data densely gather at the small numbers of SNPs and quickly become sparser as the number of SNPs is increasing. In attempt to solve the problem, GSA-SNP2 suggests another adaptively sampling approach which samples more at crowded area and less at sparse area. Assume that data D is defined by $\{d_i = (g_i, n_i)\}$, where g_i is gene score and n_i is the number of SNPs. Sampling will be made when $n_i = n_k = 2^k - 1$ where k is a natural counting parameter in range $[0, k_{max})$, and k_{max} is the minimum number making $2^{k_{max}} - 1 \geq \max(n_i)$. As a consequence, sampled gene score is defined as:

$$g_k = \frac{1}{N_k} \sum_{i=1}^{N_k} g_i$$

, where N_k is the number of data instances whose $n_i = n_k$ and g_i is the corresponding gene scores. Sampled data instances (g_k, n_k) can be used as active knots feeding into the dual cubic spline estimation algorithm.

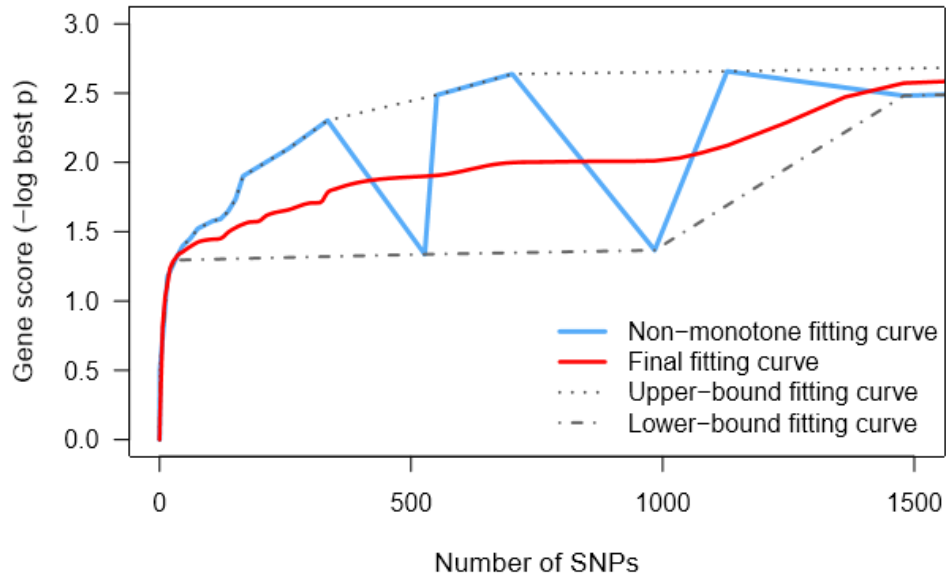


Figure S3.1. Dual cubic spline illustration.

Blue fitting curve, which is severely fluctuated, is obtained when directly applying cubic spline fitting on all data points (knots) at the same time. Dot and dash-dot fitting curves are obtained when respectively applying cubic spline fitting on upper-bound and lower-bound monotone groups of knots. Red fitting curve is the final result when applying cubic spline fitting on inferred average knots of both upper and lower curves.

15 biological processes related to height regulation

1) Skeletal system development

The abnormal skeletal system development leads to skeletal dysplasia which is the disorders of cartilage and bone. Currently, more than 350 skeletal dysplasia have been discovered caused by different types of genetic mutations ¹⁵³.

2) Epigenetics

The epigenetic regulation of height-associated genes is important for developmental process ¹⁵⁴. For example, the defects in the genomic imprinting leads to growth disorder including Silver-Russel and Beckwith-Wiedemann syndromes ^{139, 155}. Furthermore, some height-associated genes such as DOT1-like and NSD1 histone methyltransferases, HMGA1, HMGA2 are involved in the assembly of chromatin structure. Among these, the malfunction of the histone methyltransferase NSD1 causes the Sotos syndrome, characterized by overgrowth in childhood and retardation in mental and movement abilities ¹⁵⁶.

3) Insulin-like growth factor-1 and growth hormone

Insulin-like growth factor-1 mediates the protein anabolic and linear growth prompting effect of pituitary growth hormone (GH) ¹⁵⁷.

4) Wnt/ β -catenin signaling

Wnt/ β -catenin signaling affects to the skeletal development. In the early stage of skeletal development, this signaling leads mesenchymal progenitor cells to bone-forming osteoblast lineage. Later, Wnt/ β -catenin in the chondrocyte of growth plate promotes the chondrocyte survival, hypertrophic differentiation and endochondrial ossification. Functional study revealed that the mutation in Wnt signaling component affected to the bone mass in mice and human ¹⁵⁸.

5) TGF β signaling

Transforming growth factor- β (TGF β) signaling is important in chondrogenesis and osteogenesis. The defects in TGF β leads to chondrodysplasias characterized by short stature with short limbs ¹⁵⁹.

6) Platelet-derived growth factor

Platelet-derived growth factor (PDGF)-BB was reported as potent stimulator of proliferation of growth plate chondrocyte ¹⁶⁰.

7) Extracellular matrix

Growth plate is composed of ordered zone of chondrocyte which secretes extracellular matrix (ECM) including several types of collagens and proteoglycan. The mutation in ECM contributes to the abnormal growth plate development ¹⁶¹.

8) Nuclear matrix

Nuclear matrix protein Satb2 represses Hoxa2 expression and acts with other regulatory proteins to promote osteoblast differentiation ¹⁶².

9) Cell cycle

Cell proliferation is important for the normal development of mammals because their body size is predominantly determined by the number of cells. The mutations of several genes involved in DNA repair and replication cause the growth failure in primordial dwarfism ¹³⁸.

10) Androgen

Androgen secretion increases during the puberty and it regulates the rate of height growth during the adolescence, particularly in males ¹⁶³.

11) Ubiquitin ligase

E3 ubiquitin ligases c-Cbl and Cbl-b have been reported to interact with receptor tyrosine kinases (RTK) and other molecules to control the bone cell proliferation, differentiation and survival. The inhibition of c-Cbl promotes the osteoblast differentiation through the decreased RTK degradation ¹⁶⁴.

12) Nuclear Factor Kappa B

Nuclear factor kappa B (NF-kB) is expressed in the growth plate chondrocyte and it mediates the promoting effect of growth hormone and IGF-1 on longitudinal bone growth and growth plate chondrogenesis ¹⁶⁵.

13) Proteoglycan

The proteoglycans are components in the extracellular matrix of cartilage. It is essential during the differentiation and for maintenance of cartilaginous skeletal elements ¹⁴¹.

14) Reactive oxygen species

Reactive oxygen species (ROS) are important components that regulate the differentiation and bone-resorbing function of osteoclast ¹⁴².

15) Fibroblast growth factor and telomerase activity

Fibroblast growth factor receptor 3 (FGFR3) inhibits chondrocyte proliferation by down-regulating the telomerase reverse transcriptase (TERT) and reducing the telomerase activity. It suggests the important role of telomerase activity in the chondrocyte proliferation during the bone elongation¹⁴³.

Table S3.1. Gene Ontology terms (mSigDB C5 v6.0) related to 15 height-related categories

Skeletal system development	
GO APPENDAGE DEVELOPMENT	GO NEGATIVE REGULATION OF OSSIFICATION
GO BONE CELL DEVELOPMENT	GO NEGATIVE REGULATION OF OSTEOBLAST DIFFERENTIATION
GO BONE DEVELOPMENT	GO NEGATIVE REGULATION OF SKELETAL MUSCLE TISSUE DEVELOPMENT
GO BONE GROWTH	GO OSSIFICATION
GO BONE MATURATION	GO OSTEOBLAST DEVELOPMENT
GO BONE MINERALIZATION	GO OSTEOBLAST DIFFERENTIATION
GO BONE MORPHOGENESIS	GO PARAXIAL MESODERM DEVELOPMENT
GO BONE REMODELING	GO POSITIVE REGULATION OF BONE REMODELING
GO BONE RESORPTION	GO POSITIVE REGULATION OF CARTILAGE DEVELOPMENT
GO BONE TRABECULA MORPHOGENESIS	GO POSITIVE REGULATION OF CHONDROCYTE DIFFERENTIATION
GO CARTILAGE DEVELOPMENT	GO POSITIVE REGULATION OF OSSIFICATION
GO CARTILAGE DEVELOPMENT INVOLVED IN ENDOCHONDRAL BONE MORPHOGENESIS	GO POSITIVE REGULATION OF OSTEOBLAST DIFFERENTIATION
GO CARTILAGE MORPHOGENESIS	GO POSITIVE REGULATION OF OSTEOBLAST PROLIFERATION
GO CHONDROCYTE DEVELOPMENT	GO POSITIVE REGULATION OF SKELETAL MUSCLE TISSUE DEVELOPMENT
GO CHONDROCYTE DIFFERENTIATION	GO REGULATION OF BONE DEVELOPMENT
GO CONNECTIVE TISSUE DEVELOPMENT	GO REGULATION OF BONE REMODELING
GO CRANIAL SKELETAL SYSTEM DEVELOPMENT	GO REGULATION OF BONE RESORPTION
GO EMBRYONIC CRANIAL SKELETON MORPHOGENESIS	GO REGULATION OF CARTILAGE DEVELOPMENT
GO EMBRYONIC FORELIMB MORPHOGENESIS	GO REGULATION OF CHONDROCYTE DIFFERENTIATION
GO EMBRYONIC HINDLIMB MORPHOGENESIS	GO REGULATION OF MESODERM DEVELOPMENT
GO EMBRYONIC SKELETAL JOINT DEVELOPMENT	GO REGULATION OF OSSIFICATION
GO EMBRYONIC SKELETAL SYSTEM DEVELOPMENT	GO REGULATION OF OSTEOBLAST DIFFERENTIATION
GO EMBRYONIC SKELETAL SYSTEM MORPHOGENESIS	GO REGULATION OF OSTEOBLAST PROLIFERATION
GO ENDOCHONDRAL BONE MORPHOGENESIS	GO REGULATION OF SKELETAL MUSCLE ADAPTATION
GO FORELIMB MORPHOGENESIS	GO REGULATION OF SKELETAL MUSCLE CELL DIFFERENTIATION
GO GROWTH PLATE CARTILAGE DEVELOPMENT	GO REGULATION OF SKELETAL MUSCLE CELL PROLIFERATION
GO HINDLIMB MORPHOGENESIS	GO REGULATION OF SKELETAL MUSCLE CONTRACTION
GO LATERAL MESODERM DEVELOPMENT	GO REGULATION OF SKELETAL MUSCLE TISSUE DEVELOPMENT
GO LIMBIC SYSTEM DEVELOPMENT	GO REPLACEMENT OSSIFICATION
GO MESODERM DEVELOPMENT	GO SKELETAL MUSCLE ADAPTATION
GO MESODERM MORPHOGENESIS	GO SKELETAL MUSCLE CELL DIFFERENTIATION
GO MESODERMAL CELL DIFFERENTIATION	GO SKELETAL MUSCLE CONTRACTION
GO MESODERMAL CELL FATE COMMITMENT	GO SKELETAL MUSCLE ORGAN DEVELOPMENT
GO NEGATIVE REGULATION OF BONE REMODELING	GO SKELETAL MUSCLE TISSUE REGENERATION
GO NEGATIVE REGULATION OF BONE RESORPTION	GO SKELETAL SYSTEM DEVELOPMENT
GO NEGATIVE REGULATION OF CARTILAGE DEVELOPMENT	GO SKELETAL SYSTEM MORPHOGENESIS
GO NEGATIVE REGULATION OF CHONDROCYTE DIFFERENTIATION	
Epigenetics	
GO ATP DEPENDENT CHROMATIN REMODELING	GO HISTONE METHYLTRANSFERASE ACTIVITY H3 K4 SPECIFIC
GO CHROMATIN	GO HISTONE METHYLTRANSFERASE COMPLEX
GO CHROMATIN ASSEMBLY OR DISASSEMBLY	GO HISTONE MONOUBIQUITINATION
GO CHROMATIN BINDING	GO HISTONE MRNA CATABOLIC PROCESS

GO CHROMATIN DISASSEMBLY	GO HISTONE MRNA METABOLIC PROCESS
GO CHROMATIN DNA BINDING	GO HISTONE PHOSPHORYLATION
GO CHROMATIN MODIFICATION	GO HISTONE UBIQUITINATION
GO CHROMATIN ORGANIZATION	GO LYSINE ACETYLATED HISTONE BINDING
GO CHROMATIN REMODELING	GO METHYLATED HISTONE BINDING
GO CHROMATIN SILENCING	GO METHYLATION DEPENDENT CHROMATIN SILENCING
GO CHROMATIN SILENCING AT RDNA	GO NEGATIVE REGULATION OF CHROMATIN MODIFICATION
GO COVALENT CHROMATIN MODIFICATION	GO NEGATIVE REGULATION OF GENE EXPRESSION EPIGENETIC
GO DNA PACKAGING	GO NEGATIVE REGULATION OF GENE SILENCING
GO DNA PACKAGING COMPLEX	GO NEGATIVE REGULATION OF HISTONE ACETYLATION
GO DNA REPLICATION DEPENDENT NUCLEOSOME ORGANIZATION	GO NEGATIVE REGULATION OF HISTONE METHYLATION
GO DNA REPLICATION INDEPENDENT NUCLEOSOME ORGANIZATION	GO NEGATIVE REGULATION OF HISTONE MODIFICATION
GO EUCHROMATIN	GO NUCLEAR CHROMATIN
GO GENE SILENCING	GO NUCLEAR EUCHROMATIN
GO H4 HISTONE ACETYLTRANSFERASE ACTIVITY	GO NUCLEAR HETEROCHROMATIN
GO H4 HISTONE ACETYLTRANSFERASE COMPLEX	GO NUCLEAR NUCLEOSOME
GO HETEROCHROMATIN	GO NUCLEOSOME BINDING
GO HETEROCHROMATIN ORGANIZATION	GO PERICENTRIC HETEROCHROMATIN
GO HISTONE ACETYLTRANSFERASE BINDING	GO POSITIVE REGULATION OF CHROMATIN MODIFICATION
GO HISTONE BINDING	GO POSITIVE REGULATION OF GENE EXPRESSION EPIGENETIC
GO HISTONE DEACETYLASE ACTIVITY H3 K14 SPECIFIC	GO POSITIVE REGULATION OF HISTONE DEACETYLATION
GO HISTONE DEACETYLASE BINDING	GO POSITIVE REGULATION OF HISTONE H3 K4 METHYLATION
GO HISTONE DEACETYLASE COMPLEX	GO POSITIVE REGULATION OF HISTONE METHYLATION
GO HISTONE DEMETHYLASE ACTIVITY	GO POSTTRANSCRIPTIONAL GENE SILENCING
GO HISTONE DEUBIQUITINATION	GO PROTEIN HETEROTETRAMERIZATION
GO HISTONE EXCHANGE	GO PROTEIN LOCALIZATION TO CHROMATIN
GO HISTONE H2A ACETYLATION	GO REGULATION OF CHROMATIN BINDING
GO HISTONE H2A MONOUBIQUITINATION	GO REGULATION OF CHROMATIN ORGANIZATION
GO HISTONE H2A UBIQUITINATION	GO REGULATION OF CHROMATIN SILENCING
GO HISTONE H3 ACETYLATION	GO REGULATION OF GENE EXPRESSION EPIGENETIC
GO HISTONE H3 DEACETYLATION	GO REGULATION OF GENE SILENCING
GO HISTONE H3 K4 METHYLATION	GO REGULATION OF HISTONE DEACETYLATION
GO HISTONE H3 K4 TRIMETHYLATION	GO REGULATION OF HISTONE H3 K4 METHYLATION
GO HISTONE H3 K9 MODIFICATION	GO REGULATION OF HISTONE H3 K9 ACETYLATION
GO HISTONE H4 ACETYLATION	GO REGULATION OF HISTONE H3 K9 METHYLATION
GO HISTONE H4 K16 ACETYLATION	GO REGULATION OF HISTONE H4 ACETYLATION
GO HISTONE KINASE ACTIVITY	GO REGULATION OF HISTONE METHYLATION
GO HISTONE LYSINE N METHYLTRANSFERASE ACTIVITY	GO REGULATION OF HISTONE PHOSPHORYLATION
GO HISTONE METHYLATION	GO REGULATION OF POSTTRANSCRIPTIONAL GENE SILENCING
GO HISTONE METHYLTRANSFERASE ACTIVITY	GO TRANSCRIPTIONALLY ACTIVE CHROMATIN
Insulin-like growth factor-1 and growth hormone	
GO CELLULAR RESPONSE TO GROWTH HORMONE STIMULUS	GO POSITIVE REGULATION OF INSULIN LIKE GROWTH FACTOR RECEPTOR SIGNALING PATHWAY
GO INSULIN LIKE GROWTH FACTOR BINDING	GO REGULATION OF GROWTH HORMONE SECRETION

GO INSULIN LIKE GROWTH FACTOR RECEPTOR BINDING	GO REGULATION OF INSULIN LIKE GROWTH FACTOR RECEPTOR SIGNALING PATHWAY
GO INSULIN LIKE GROWTH FACTOR RECEPTOR SIGNALING PATHWAY	GO RESPONSE TO GROWTH HORMONE
GO JAK STAT CASCADE INVOLVED IN GROWTH HORMONE SIGNALING PATHWAY	
Wnt/β-catenin signaling	
GO BETA CATENIN BINDING	GO POSITIVE REGULATION OF WNT SIGNALING PATHWAY
GO BETA CATENIN DESTRUCTION COMPLEX	GO REGULATION OF CANONICAL WNT SIGNALING PATHWAY
GO BETA CATENIN DESTRUCTION COMPLEX DISASSEMBLY	GO REGULATION OF NON CANONICAL WNT SIGNALING PATHWAY
GO BETA CATENIN TCF COMPLEX ASSEMBLY	GO REGULATION OF WNT SIGNALING PATHWAY
GO CANONICAL WNT SIGNALING PATHWAY	GO REGULATION OF WNT SIGNALING PATHWAY PLANAR CELL POLARITY PATHWAY
GO NEGATIVE REGULATION OF CANONICAL WNT SIGNALING PATHWAY	GO WNT ACTIVATED RECEPTOR ACTIVITY
GO NEGATIVE REGULATION OF WNT SIGNALING PATHWAY	GO WNT PROTEIN BINDING
GO NON-CANONICAL WNT SIGNALING PATHWAY	GO WNT SIGNALING PATHWAY
GO POSITIVE REGULATION OF CANONICAL WNT SIGNALING PATHWAY	GO WNT SIGNALING PATHWAY CALCIUM MODULATING PATHWAY
GO POSITIVE REGULATION OF NON-CANONICAL WNT SIGNALING PATHWAY	GO WNT SIGNALOSOME
TGFβ signaling	
GO NEGATIVE REGULATION OF TRANSFORMING GROWTH FACTOR BETA RECEPTOR SIGNALING PATHWAY	GO RESPONSE TO TRANSFORMING GROWTH FACTOR BETA
GO POSITIVE REGULATION OF CELLULAR RESPONSE TO TRANSFORMING GROWTH FACTOR BETA STIMULUS	GO TRANSFORMING GROWTH FACTOR BETA BINDING
GO POSITIVE REGULATION OF TRANSFORMING GROWTH FACTOR BETA PRODUCTION	GO TRANSFORMING GROWTH FACTOR BETA RECEPTOR BINDING
GO REGULATION OF CELLULAR RESPONSE TO TRANSFORMING GROWTH FACTOR BETA STIMULUS	GO TRANSFORMING GROWTH FACTOR BETA RECEPTOR SIGNALING PATHWAY
GO REGULATION OF TRANSFORMING GROWTH FACTOR BETA PRODUCTION	
Platelet-derived growth factor	
GO PLATELET DERIVED GROWTH FACTOR BINDING	GO REGULATION OF PLATELET DERIVED GROWTH FACTOR RECEPTOR SIGNALING PATHWAY
GO PLATELET DERIVED GROWTH FACTOR RECEPTOR BINDING	GO RESPONSE TO PLATELET DERIVED GROWTH FACTOR
GO PLATELET DERIVED GROWTH FACTOR RECEPTOR SIGNALING PATHWAY	
Extracellular matrix	
GO BANDED COLLAGEN FIBRIL	GO EXTRACELLULAR MATRIX DISASSEMBLY
GO BASEMENT MEMBRANE	GO EXTRACELLULAR MATRIX STRUCTURAL CONSTITUENT
GO BASEMENT MEMBRANE ORGANIZATION	GO HEPARAN SULFATE PROTEOGLYCAN BINDING
GO CHONDROITIN SULFATE PROTEOGLYCAN BIOSYNTHETIC PROCESS	GO HEPARAN SULFATE PROTEOGLYCAN BIOSYNTHETIC PROCESS
GO CHONDROITIN SULFATE PROTEOGLYCAN METABOLIC PROCESS	GO HEPARAN SULFATE PROTEOGLYCAN METABOLIC PROCESS
GO COLLAGEN BINDING	GO POSITIVE REGULATION OF EXTRACELLULAR MATRIX ORGANIZATION
GO COLLAGEN FIBRIL ORGANIZATION	GO PROTEINACEOUS EXTRACELLULAR MATRIX
GO COLLAGEN TRIMER	GO PROTEOGLYCAN BINDING
GO COMPLEX OF COLLAGEN TRIMERS	GO PROTEOGLYCAN BIOSYNTHETIC PROCESS
GO DERMATAN SULFATE PROTEOGLYCAN METABOLIC PROCESS	GO PROTEOGLYCAN METABOLIC PROCESS
GO EXTRACELLULAR MATRIX	GO REGULATION OF EXTRACELLULAR MATRIX ASSEMBLY
GO EXTRACELLULAR MATRIX ASSEMBLY	GO REGULATION OF EXTRACELLULAR MATRIX DISASSEMBLY
GO EXTRACELLULAR MATRIX BINDING	GO REGULATION OF EXTRACELLULAR MATRIX ORGANIZATION

GO EXTRACELLULAR MATRIX COMPONENT	
Nuclear matrix	
GO NUCLEAR MATRIX	
Cell cycle	
GO CELL CYCLE	GO NEGATIVE REGULATION OF CELL CYCLE ARREST
GO CELL CYCLE ARREST	GO NEGATIVE REGULATION OF CELL CYCLE G1 S PHASE TRANSITION
GO CELL CYCLE CHECKPOINT	GO NEGATIVE REGULATION OF CELL CYCLE G2 M PHASE TRANSITION
GO CELL CYCLE DNA REPLICATION	GO NEGATIVE REGULATION OF CELL CYCLE PHASE TRANSITION
GO CELL CYCLE G1 S PHASE TRANSITION	GO NEGATIVE REGULATION OF CELL CYCLE PROCESS
GO CELL CYCLE G2 M PHASE TRANSITION	GO NEGATIVE REGULATION OF DNA DEPENDENT DNA REPLICATION
GO CELL CYCLE PHASE TRANSITION	GO NEGATIVE REGULATION OF DNA REPLICATION
GO CELL CYCLE PROCESS	GO NEGATIVE REGULATION OF MEIOTIC CELL CYCLE
GO CHROMOSOME ORGANIZATION INVOLVED IN MEIOTIC CELL CYCLE	GO NEGATIVE REGULATION OF MITOTIC CELL CYCLE
GO DNA DEPENDENT DNA REPLICATION	GO NEGATIVE REGULATION OF MITOTIC NUCLEAR DIVISION
GO DNA DEPENDENT DNA REPLICATION MAINTENANCE OF FIDELITY	GO POSITIVE REGULATION OF CELL CYCLE
GO DNA REPLICATION	GO POSITIVE REGULATION OF CELL CYCLE ARREST
GO DNA REPLICATION CHECKPOINT	GO POSITIVE REGULATION OF CELL CYCLE G1 S PHASE TRANSITION
GO DNA REPLICATION DEPENDENT NUCLEOSOME ORGANIZATION	GO POSITIVE REGULATION OF CELL CYCLE G2 M PHASE TRANSITION
GO DNA REPLICATION FACTOR A COMPLEX	GO POSITIVE REGULATION OF CELL CYCLE PHASE TRANSITION
GO DNA REPLICATION INDEPENDENT NUCLEOSOME ORGANIZATION	GO POSITIVE REGULATION OF CELL CYCLE PROCESS
GO DNA REPLICATION INITIATION	GO POSITIVE REGULATION OF DNA DEPENDENT DNA REPLICATION
GO DNA STRAND ELONGATION INVOLVED IN DNA REPLICATION	GO POSITIVE REGULATION OF DNA REPLICATION
GO ESTABLISHMENT OF MITOTIC SPINDLE LOCALIZATION	GO POSITIVE REGULATION OF G1 S TRANSITION OF MITOTIC CELL CYCLE
GO ESTABLISHMENT OF MITOTIC SPINDLE ORIENTATION	GO POSITIVE REGULATION OF MEIOTIC CELL CYCLE
GO MEIOTIC CELL CYCLE	GO POSITIVE REGULATION OF MITOTIC CELL CYCLE
GO MEIOTIC CELL CYCLE PROCESS	GO POSITIVE REGULATION OF MITOTIC NUCLEAR DIVISION
GO MITOTIC CELL CYCLE	GO POSITIVE REGULATION OF MITOTIC SISTER CHROMATID SEPARATION
GO MITOTIC CELL CYCLE ARREST	GO REGULATION OF CELL CYCLE
GO MITOTIC CELL CYCLE CHECKPOINT	GO REGULATION OF CELL CYCLE ARREST
GO MITOTIC CHROMOSOME CONDENSATION	GO REGULATION OF CELL CYCLE CHECKPOINT
GO MITOTIC CYTOKINESIS	GO REGULATION OF CELL CYCLE G1 S PHASE TRANSITION
GO MITOTIC DNA INTEGRITY CHECKPOINT	GO REGULATION OF CELL CYCLE G2 M PHASE TRANSITION
GO MITOTIC G2 DNA DAMAGE CHECKPOINT	GO REGULATION OF CELL CYCLE PHASE TRANSITION
GO MITOTIC G2 M TRANSITION CHECKPOINT	GO REGULATION OF CELL CYCLE PROCESS
GO MITOTIC NUCLEAR DIVISION	GO REGULATION OF DNA DEPENDENT DNA REPLICATION
GO MITOTIC RECOMBINATION	GO REGULATION OF DNA REPLICATION
GO MITOTIC SISTER CHROMATID COHESION	GO REGULATION OF MEIOTIC CELL CYCLE
GO MITOTIC SISTER CHROMATID SEGREGATION	GO REGULATION OF MITOTIC CELL CYCLE
GO MITOTIC SPINDLE	GO REGULATION OF MITOTIC SPINDLE CHECKPOINT
GO MITOTIC SPINDLE ASSEMBLY	GO REGULATION OF NUCLEAR CELL CYCLE DNA REPLICATION
GO MITOTIC SPINDLE ORGANIZATION	GO REGULATION OF TRANSCRIPTION INVOLVED IN G1 S TRANSITION OF MITOTIC CELL CYCLE

GO NEGATIVE REGULATION OF CELL CYCLE	
Androgen	
GO ANDROGEN BIOSYNTHETIC PROCESS	GO ANDROGEN RECEPTOR SIGNALING PATHWAY
GO ANDROGEN METABOLIC PROCESS	GO NEGATIVE REGULATION OF ANDROGEN RECEPTOR SIGNALING PATHWAY
GO ANDROGEN RECEPTOR BINDING	GO REGULATION OF ANDROGEN RECEPTOR SIGNALING PATHWAY
Proteoglycan	
GO CHONDROITIN SULFATE PROTEOGLYCAN BIOSYNTHETIC PROCESS	GO HEPARAN SULFATE PROTEOGLYCAN BIOSYNTHETIC PROCESS
GO CHONDROITIN SULFATE PROTEOGLYCAN METABOLIC PROCESS	GO HEPARAN SULFATE PROTEOGLYCAN METABOLIC PROCESS
GO DERMATAN SULFATE PROTEOGLYCAN METABOLIC PROCESS	GO PROTEOGLYCAN BINDING
GO GLYCOSAMINOGLYCAN BINDING	GO PROTEOGLYCAN BIOSYNTHETIC PROCESS
GO HEPARAN SULFATE PROTEOGLYCAN BINDING	GO PROTEOGLYCAN METABOLIC PROCESS
Reactive oxygen species	
GO CELL DEATH IN RESPONSE TO OXIDATIVE STRESS	GO REACTIVE OXYGEN SPECIES BIOSYNTHETIC PROCESS
GO CELLULAR RESPONSE TO OXIDATIVE STRESS	GO REACTIVE OXYGEN SPECIES METABOLIC PROCESS
GO CELLULAR RESPONSE TO REACTIVE OXYGEN SPECIES	GO REGULATION OF OXIDATIVE STRESS INDUCED CELL DEATH
GO NEGATIVE REGULATION OF OXIDATIVE STRESS INDUCED INTRINSIC APOPTOTIC SIGNALING PATHWAY	GO REGULATION OF OXIDATIVE STRESS INDUCED INTRINSIC APOPTOTIC SIGNALING PATHWAY
GO NEGATIVE REGULATION OF REACTIVE OXYGEN SPECIES BIOSYNTHETIC PROCESS	GO REGULATION OF OXIDATIVE STRESS INDUCED NEURON DEATH
GO NEGATIVE REGULATION OF REACTIVE OXYGEN SPECIES METABOLIC PROCESS	GO REGULATION OF REACTIVE OXYGEN SPECIES BIOSYNTHETIC PROCESS
GO NEGATIVE REGULATION OF RESPONSE TO OXIDATIVE STRESS	GO REGULATION OF REACTIVE OXYGEN SPECIES METABOLIC PROCESS
GO NEGATIVE REGULATION OF RESPONSE TO REACTIVE OXYGEN SPECIES	GO REGULATION OF RESPONSE TO OXIDATIVE STRESS
GO POSITIVE REGULATION OF OXIDATIVE STRESS INDUCED CELL DEATH	GO REGULATION OF RESPONSE TO REACTIVE OXYGEN SPECIES
GO POSITIVE REGULATION OF REACTIVE OXYGEN SPECIES BIOSYNTHETIC PROCESS	GO RESPONSE TO OXIDATIVE STRESS
GO POSITIVE REGULATION OF REACTIVE OXYGEN SPECIES METABOLIC PROCESS	GO RESPONSE TO REACTIVE OXYGEN SPECIES
GO POSITIVE REGULATION OF RESPONSE TO OXIDATIVE STRESS	
Fibroblast growth factor and telomerase activity	
GO CHROMOSOME TELOMERIC REGION	GO REGULATION OF TELOMERASE ACTIVITY
GO FIBROBLAST GROWTH FACTOR BINDING	GO REGULATION OF TELOMERASE RNA LOCALIZATION TO CAJAL BODY
GO FIBROBLAST GROWTH FACTOR RECEPTOR BINDING	GO REGULATION OF TELOMERE CAPPING
GO FIBROBLAST GROWTH FACTOR RECEPTOR SIGNALING PATHWAY	GO REGULATION OF TELOMERE MAINTENANCE
GO NEGATIVE REGULATION OF FIBROBLAST GROWTH FACTOR RECEPTOR SIGNALING PATHWAY	GO REGULATION OF TELOMERE MAINTENANCE VIA TELOMERE LENGTHENING
GO NEGATIVE REGULATION OF TELOMERASE ACTIVITY	GO RESPONSE TO FIBROBLAST GROWTH FACTOR
GO NEGATIVE REGULATION OF TELOMERE MAINTENANCE	GO TELOMERASE HOLOENZYME COMPLEX
GO NEGATIVE REGULATION OF TELOMERE MAINTENANCE VIA TELOMERASE	GO TELOMERASE RNA BINDING
GO NEGATIVE REGULATION OF TELOMERE MAINTENANCE VIA TELOMERE LENGTHENING	GO TELOMERE CAP COMPLEX
GO NUCLEAR CHROMOSOME TELOMERIC REGION	GO TELOMERE CAPPING
GO POSITIVE REGULATION OF TELOMERASE ACTIVITY	GO TELOMERE LOCALIZATION
GO POSITIVE REGULATION OF TELOMERE CAPPING	GO TELOMERE MAINTENANCE VIA RECOMBINATION
GO POSITIVE REGULATION OF TELOMERE MAINTENANCE	GO TELOMERE MAINTENANCE VIA TELOMERASE
GO POSITIVE REGULATION OF TELOMERE MAINTENANCE VIA TELOMERE LENGTHENING	GO TELOMERE MAINTENANCE VIA TELOMERE LENGTHENING

GO PROTEIN LOCALIZATION TO CHROMOSOME TELOMERIC REGION	GO TELOMERE ORGANIZATION
GO REGULATION OF FIBROBLAST GROWTH FACTOR RECEPTOR SIGNALING PATHWAY	GO TELOMERIC DNA BINDING
GO REGULATION OF PROTEIN LOCALIZATION TO CHROMOSOME TELOMERIC REGION	
Ubiquitin ligase	
GO CUL3 RING UBIQUITIN LIGASE COMPLEX	GO PROTEIN UBIQUITINATION
GO CUL4 RING E3 UBIQUITIN LIGASE COMPLEX	GO PROTEIN UBIQUITINATION INVOLVED IN UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS
GO CULLIN RING UBIQUITIN LIGASE COMPLEX	GO REGULATION OF PROTEASOMAL UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS
GO CYTOPLASMIC UBIQUITIN LIGASE COMPLEX	GO REGULATION OF PROTEIN POLYUBIQUITINATION
GO ER ASSOCIATED UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS	GO REGULATION OF PROTEIN UBIQUITINATION INVOLVED IN UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS
GO HISTONE DEUBIQUITINATION	GO REGULATION OF UBIQUITIN PROTEIN LIGASE ACTIVITY
GO HISTONE H2A MONOUBIQUITINATION	GO SCF DEPENDENT PROTEASOMAL UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS
GO HISTONE H2A UBIQUITINATION	GO SCF UBIQUITIN LIGASE COMPLEX
GO HISTONE MONOUBIQUITINATION	GO THIOL DEPENDENT UBIQUITIN SPECIFIC PROTEASE ACTIVITY
GO HISTONE UBIQUITINATION	GO UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS VIA THE MULTIVESICULAR BODY SORTING PATHWAY
GO K63 LINKED POLYUBIQUITIN BINDING	GO UBIQUITIN LIGASE COMPLEX
GO NUCLEAR UBIQUITIN LIGASE COMPLEX	GO UBIQUITIN LIKE PROTEIN BINDING
GO POLYUBIQUITIN BINDING	GO UBIQUITIN LIKE PROTEIN CONJUGATING ENZYME ACTIVITY
GO PROTEIN AUTOUBIQUITINATION	GO UBIQUITIN LIKE PROTEIN CONJUGATING ENZYME BINDING
GO PROTEIN K11 LINKED UBIQUITINATION	GO UBIQUITIN LIKE PROTEIN LIGASE ACTIVITY
GO PROTEIN K48 LINKED DEUBIQUITINATION	GO UBIQUITIN LIKE PROTEIN LIGASE BINDING
GO PROTEIN K48 LINKED UBIQUITINATION	GO UBIQUITIN LIKE PROTEIN SPECIFIC PROTEASE ACTIVITY
GO PROTEIN K63 LINKED DEUBIQUITINATION	GO UBIQUITIN LIKE PROTEIN TRANSFERASE ACTIVITY
GO PROTEIN K63 LINKED UBIQUITINATION	GO UBIQUITIN SPECIFIC PROTEASE BINDING
GO PROTEIN MONOUBIQUITINATION	GO UBIQUITIN UBIQUITIN LIGASE ACTIVITY
GO PROTEIN POLYUBIQUITINATION	
Nuclear Factor Kappa B	
GO ACTIVATION OF NF KAPPAB INDUCING KINASE ACTIVITY	GO POSITIVE REGULATION OF I KAPPAB KINASE NF KAPPAB SIGNALING
GO I KAPPAB KINASE NF KAPPAB SIGNALING	GO POSITIVE REGULATION OF NF KAPPAB IMPORT INTO NUCLEUS
GO NEGATIVE REGULATION OF I KAPPAB KINASE NF KAPPAB SIGNALING	GO POSITIVE REGULATION OF NF KAPPAB TRANSCRIPTION FACTOR ACTIVITY
GO NEGATIVE REGULATION OF NF KAPPAB IMPORT INTO NUCLEUS	GO POSITIVE REGULATION OF NIK NF KAPPAB SIGNALING
GO NEGATIVE REGULATION OF NF KAPPAB TRANSCRIPTION FACTOR ACTIVITY	GO REGULATION OF I KAPPAB KINASE NF KAPPAB SIGNALING
GO NF KAPPAB BINDING	GO REGULATION OF NF KAPPAB IMPORT INTO NUCLEUS
GO NIK NF KAPPAB SIGNALING	GO REGULATION OF NIK NF KAPPAB SIGNALING

Table S3.2. Canonical pathways (mSigDB C2 v6.0) related to 15 T2D-related pathways

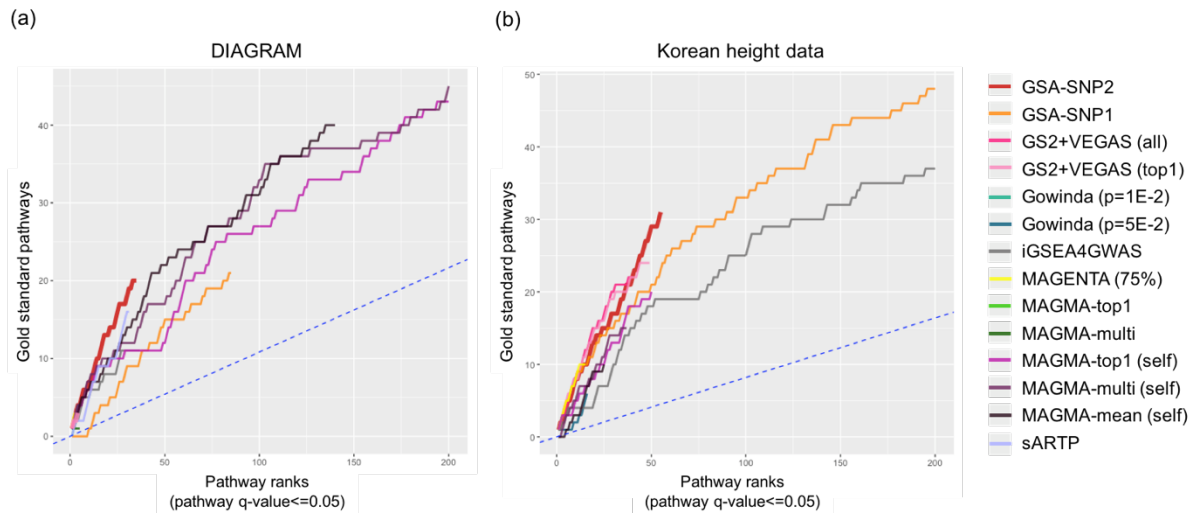
Diabetes	
KEGG MATURITY ONSET DIABETES OF THE YOUNG	KEGG TYPE II DIABETES MELLITUS
KEGG TYPE I DIABETES MELLITUS	REACTOME DIABETES PATHWAYS
Regulation of beta cell	
REACTOME REGULATION OF BETA CELL DEVELOPMENT	PID HNF3B PATHWAY
REACTOME REGULATION OF GENE EXPRESSION IN BETA CELLS	
Insulin/blood glucose level	
REACTOME REGULATION OF INSULIN SECRETION	REACTOME GLUCAGON SIGNALING IN METABOLIC REGULATION
REACTOME REGULATION OF INSULIN SECRETION BY GLUCAGON LIKE PEPTIDE1	REACTOME POTASSIUM CHANNELS
KEGG INSULIN SIGNALING PATHWAY	REACTOME VOLTAGE GATED POTASSIUM CHANNELS
SIG INSULIN RECEPTOR PATHWAY IN CARDIAC MYOCYTES	REACTOME ION CHANNEL TRANSPORT
REACTOME INSULIN RECEPTOR RECYCLING	REACTOME INWARDLY RECTIFYING K CHANNELS
REACTOME REGULATION OF INSULIN SECRETION BY ACETYLCHOLINE	REACTOME TANDEM PORE DOMAIN POTASSIUM CHANNELS
REACTOME INHIBITION OF INSULIN SECRETION BY ADRENALINE NORADRENALINE	REACTOME INHIBITION OF VOLTAGE GATED CA2 CHANNELS VIA GBETA GAMMA SUBUNITS
PID INSULIN GLUCOSE PATHWAY	REACTOME ADENYLATE CYCLASE ACTIVATING PATHWAY
BIOCARTA INSULIN PATHWAY	REACTOME ADENYLATE CYCLASE INHIBITORY PATHWAY
REACTOME SIGNALING BY INSULIN RECEPTOR	REACTOME GLUCAGON TYPE LIGAND RECEPTORS
REACTOME INSULIN RECEPTOR SIGNALLING CASCADE	KEGG CALCIUM SIGNALING PATHWAY
REACTOME INSULIN SYNTHESIS AND PROCESSING	REACTOME ION TRANSPORT BY P TYPE ATPASES
PID INSULIN PATHWAY	
Adipocytokine signaling	
KEGG ADIPOCYTOKINE SIGNALING PATHWAY	BIOCARTA IL6 PATHWAY
BIOCARTA LEPTIN PATHWAY	PID TNF PATHWAY
PID IL6 7 PATHWAY	ST TUMOR NECROSIS FACTOR PATHWAY
Cell cycle	
REACTOME G0 AND EARLY G1	REACTOME P53 INDEPENDENT G1 S DNA DAMAGE CHECKPOINT
KEGG CELL CYCLE	REACTOME MITOTIC M M G1 PHASES
REACTOME CELL CYCLE CHECKPOINTS	REACTOME G1 S TRANSITION
REACTOME REGULATION OF MITOTIC CELL CYCLE	REACTOME M G1 TRANSITION
BIOCARTA G1 PATHWAY	BIOCARTA G2 PATHWAY
REACTOME G1 PHASE	REACTOME MITOTIC G2 G2 M PHASES
SA G1 AND S PHASES	REACTOME CYCLIN A B1 ASSOCIATED EVENTS DURING G2 M TRANSITION
REACTOME MITOTIC G1 G1 S PHASES	REACTOME G2 M CHECKPOINTS
REACTOME G1 S SPECIFIC TRANSCRIPTION	REACTOME S PHASE
REACTOME P53 DEPENDENT G1 DNA DAMAGE RESPONSE	BIOCARTA CELLCYCLE PATHWAY
REACTOME CYCLIN E ASSOCIATED EVENTS DURING G1 S TRANSITION	REACTOME MITOTIC PROMETAPHASE
Circadian rhythm	
PID CIRCADIAN PATHWAY	REACTOME CIRCADIAN REPRESSION OF EXPRESSION BY REV ERBA
KEGG CIRCADIAN RHYTHM MAMMAL	REACTOME BMAL1 CLOCK NPAS2 ACTIVATES CIRCADIAN EXPRESSION
REACTOME RORA ACTIVATES CIRCADIAN EXPRESSION	REACTOME CIRCADIAN CLOCK
Unfolded protein response	
REACTOME UNFOLDED PROTEIN RESPONSE	REACTOME ACTIVATION OF CHAPERONE GENES BY XBP1S
Branched-chain amino acid metabolism	

KEGG VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	KEGG VALINE LEUCINE AND ISOLEUCINE DEGRADATION
Fatty acid metabolism	
KEGG BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	REACTOME FATTY ACYL COA BIOSYNTHESIS
REACTOME FATTY ACID TRIACYLGLYCEROL AND KETONE BODY METABOLISM	REACTOME MITOCHONDRIAL FATTY ACID BETA OXIDATION
REACTOME ACTIVATED AMPK STIMULATES FATTY ACID OXIDATION IN MUSCLE	REACTOME TRIGLYCERIDE BIOSYNTHESIS
KEGG FATTY ACID METABOLISM	GERHOLD ADIPOGENESIS UP
REACTOME SYNTHESIS OF VERY LONG CHAIN FATTY ACYL COAS	REACTOME TRANSCRIPTIONAL REGULATION OF WHITE ADIPOCYTE DIFFERENTIATION
Glycolysis and Gluconeogenesis	
KEGG GLYCOLYSIS GLUCONEOGENESIS	REACTOME GLUCONEOGENESIS
REACTOME GLYCOLYSIS	
Inflammation	
BIOCARTA INFLAM PATHWAY	REACTOME TAK1 ACTIVATES NFKB BY PHOSPHORYLATION AND ACTIVATION OF IKKS COMPLEX
REACTOME THE NLRP3 INFLAMMASOME	PID NFKAPPAB CANONICAL PATHWAY
REACTOME INFLAMMASOMES	PID NFKAPPAB ATYPICAL PATHWAY
PID CXCR4 PATHWAY	REACTOME TRAF6 MEDIATED INDUCTION OF NFKB AND MAP KINASES UPON TLR7 8 OR 9 ACTIVATION
BIOCARTA CXCR4 PATHWAY	BIOCARTA NFKB PATHWAY
PID CXCR3 PATHWAY	REACTOME NFKB AND MAP KINASES ACTIVATION MEDIATED BY TLR4 SIGNALING REPERTOIRE
PID IL8 CXCR2 PATHWAY	REACTOME P75NTR SIGNALS VIA NFKB
PID IL8 CXCR1 PATHWAY	PID IL12 STAT4 PATHWAY
PID AMB2 NEUTROPHILS PATHWAY	PID IL2 STAT5 PATHWAY
REACTOME RIP MEDIATED NFKB ACTIVATION VIA DAI	ST STAT3 PATHWAY
REACTOME TRAF6 MEDIATED NFKB ACTIVATION	KEGG JAK STAT SIGNALING PATHWAY REACTOME NFKB ACTIVATION THROUGH FADD RIP1 PATHWAY MEDIATED BY CASPASE 8 AND10
NOTCH signaling	
REACTOME PRE NOTCH TRANSCRIPTION AND TRANSLATION	REACTOME RECEPTOR LIGAND BINDING INITIATES THE SECOND PROTEOLYTIC CLEAVAGE OF NOTCH RECEPTOR
REACTOME NOTCH HLH TRANSCRIPTION PATHWAY	REACTOME SIGNALING BY NOTCH2
KEGG NOTCH SIGNALING PATHWAY	REACTOME ACTIVATED NOTCH1 TRANSMITS SIGNAL TO THE NUCLEUS
PID NOTCH PATHWAY	REACTOME SIGNALING BY NOTCH1
REACTOME PRE NOTCH EXPRESSION AND PROCESSING	REACTOME SIGNALING BY NOTCH3
REACTOME SIGNALING BY NOTCH	REACTOME PRE NOTCH PROCESSING IN GOLGI
REACTOME SIGNALING BY NOTCH4	REACTOME NOTCH1 INTRACELLULAR DOMAIN REGULATES TRANSCRIPTION
PPARG signaling	
KEGG PPAR SIGNALING PATHWAY	
WNT signaling	
ST WNT CA2 CYCLIC GMP PATHWAY	ST WNT BETA CATENIN PATHWAY
KEGG WNT SIGNALING PATHWAY	PID WNT SIGNALING PATHWAY
BIOCARTA WNT PATHWAY	REACTOME SIGNALING BY WNT
PID WNT NONCANONICAL PATHWAY	PID BETA CATENIN NUC PATHWAY
PID WNT CANONICAL PATHWAY	PID BETA CATENIN DEG PATHWAY
WNT SIGNALING	
Mitochondrial dysfunction	
REACTOME MITOCHONDRIAL TRNA AMINOACYLATION	REACTOME MITOCHONDRIAL PROTEIN IMPORT

REACTOME RNA POL I RNA POL III AND MITOCHONDRIAL TRANSCRIPTION	BIOCARTA MITOCHONDRIA PATHWAY
REACTOME MITOCHONDRIAL FATTY ACID BETA OXIDATION	

Figure S3.2. Power comparison using real data with strict significance cutoff.

DIAGRAM and Korean height data were re-analyzed using stricter cutoff (pathway q-value \leq 0.05). Methods that failed to detect TP terms were not represented.



Chapter IV: Biclustering analysis of transcriptome big data

identifies condition-specific miRNA targets

4.1 Abstract

Here, a novel approach was devised to identify human microRNA (miRNA) regulatory modules (mRNA targets and relevant cell conditions) by biclustering a large collection of mRNA fold-change data for sequence-specific targets. Bicluster targets were assessed using validated mRNA targets and exhibited on an average 17.0% (median 19.4%) improved gain in certainty (sensitivity + specificity). Net gain was further increased up to 32.0% (median 33.4%) by incorporating functional networks of targets. The cancer-specific biclusters were analyzed and it was found that PI3K/Akt signaling pathway was strongly enriched with targets of a few miRNAs in breast cancer and diffuse large B-cell lymphoma. Indeed, five independent prognostic miRNAs were identified, and repression of bicluster targets and pathway activity by mir-29 was experimentally validated. In total, 29,898 biclusters for 459 human miRNAs were collected in the BiMIR database where biclusters are searchable for miRNAs, tissues, diseases, keywords, and target genes.

4.2 Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules (19 - 23nt) that regulate gene expression by binding to miRNA response elements in mRNA at the post-transcription level ¹⁶⁶⁻¹⁶⁷. Since their discovery, extensive studies have revealed their key roles in regulating cell cycle and differentiation, chronic diseases, cancer progression, and other processes ¹⁶⁸⁻¹⁷¹. As the function of a miRNA is characterized by its target genes, there have been efforts to systematically identify these target genes based on the binding sequences ¹⁷²⁻¹⁷⁶. Although these methods provide hundreds to thousands of potential targets, they yield a great number of false positives and do not suggest specific targets related to the cell condition in question.

To select more reliable mRNA targets of each miRNA, expression profiles of mRNAs and miRNAs (denoted paired expression profiles) have been incorporated taking into account the anticorrelation between miRNA and its target mRNA. Besides the simple Pearson and Spearman correlation methods, a number of computational methods that integrated both the binding sequence and paired expression data have been developed to infer the miRNA-mRNA regulatory relationships including penalized regressions and Bayesian method ¹⁷⁷⁻¹⁷⁹ (denoted anticorrelation-based methods). Many of them used multivariate linear model where multiple miRNAs regulate their common target gene. These methods not only improved the target prediction but provided the cellular condition where the paired expression

data were generated. However, anticorrelation-based methods require highly costly paired expression profiles and only a limited number of such paired data are publicly available at present.

Another line of efforts to improve the miRNA target prediction was the inference of miRNA regulation modules. Based on the binding sequence information, a bipartite graph between miRNAs and mRNAs was constructed and the maximum bicliques (or biclusters) were identified¹⁸⁰⁻¹⁸¹. These bicliques represent miRNA regulation modules where multiple miRNAs may coregulate their common targets. By incorporating paired expression data, these modules were further refined for specific cell conditions¹⁸²⁻¹⁸⁵. Considering the modular nature of cellular processes, these modules were regarded to represent more reliable miRNA regulations¹⁸⁶. Recent methods incorporated additional information such as protein-protein (or gene-gene) interaction, copy number variation, as well as methylation data to better understand miRNA regulation¹⁸⁷. The myriad of computational methods for miRNA target prediction are reviewed and categorized in the literature^{179,184,187}, and some of them are summarized in Table S4.1. In this study, a novel approach was proposed to identify miRNA targets for a variety of cell conditions by biclustering a large collection of mRNA profiles for sequence-specific targets. To this end, I and three students (Hyeong Goo Kang, Jinhwan Kim, Seon-Young Hwang) collected 5,158 human microarray expression datasets with diverse test and control conditions from the Gene Expression Omnibus (GEO) database¹⁸⁸ and compiled corresponding fold-change (FC) profiles representing the 5,158 cell conditions. Whereas the existing methods for miRNA regulation modules biclustered miRNAs and mRNA targets under a given cell condition (Figure 4.1a), a different dimension that biclusters mRNA targets and cell conditions (i.e. FC profiles) for a given miRNA of interest was considered in this study (Figure 4.1b). This approach is able to provide more reliable miRNA target groups that are robustly regulated across different cell conditions without using paired expression profiles. Of course, there is a related approach that incorporated coexpression of sequence-specific targets using 250 microarray datasets to prioritize true targets¹⁸⁹, but it clustered only target genes and did not suggest relevant cell conditions.

Typically, biclustering algorithms seek to identify a complete association (namely, biclique) between two sets of nodes (e.g., set of target genes and set of cell conditions)¹⁹⁰⁻¹⁹¹. Taking into account the noise in microarray data, I developed a progressive bicluster extension (PBE) algorithm that allows for a small portion of unassociated connections between two

(a)

(b)

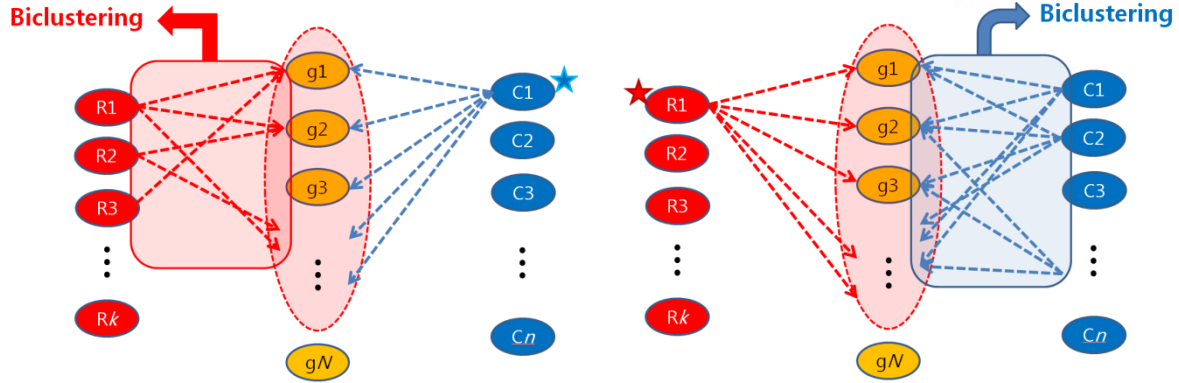


Figure 4.1. Two approaches for miRNA regulation module discovery.

Red, yellow, and blue nodes represent miRNA regulators, mRNA target genes, and cell conditions, respectively. (a) Existing approach. For a given cell condition (here, C1), down (or up)-regulated mRNAs are selected and biclusters between multiple miRNAs and these mRNA targets are sought. (b) Our approach. For a given miRNA, mRNAs with corresponding binding sequences are selected and biclusters between these mRNAs and multiple cell conditions are searched for.

node sets but yields biclusters of much larger sizes. In the initial step, PBE identifies bicliques using bimax algorithm¹⁹¹. These bicliques are used as seeds which are extended by competitively adding dense rows and columns. Then, less dense rows and columns are removed based on a threshold. By progressively applying tight to less tight thresholds during the iteration of bicluster extension, PBE was able to identify the bicluster structures from noisy data more accurately than the-state-of-the-art algorithms^{181, 191-195}. QUBIC¹⁹³ takes a similar approach that it searches for seed biclusters which are then extended. However, QUBIC only apply a threshold for minimum column density which does not change during extension.

The biclusters resulted from our method represent the miRNA target genes that show concurrent expression changes across multiple cell/tissue conditions (namely, constant biclusters). The biclusters were assessed using experimentally validated targets and exhibited substantially improved accuracy compared to the purely sequence-based method. The accuracy was even further improved by selecting the targets having functional interactions with other target genes. Notably, these gains were obtained using only publicly available gene expression and protein functional interaction data, but were compared favorably with those obtained from the anticorrelation-based methods that require costly mRNA-miRNA profiling. Moreover, our predictions are available for 459 human miRNAs and a variety of cell conditions from our bicluster database, called BiMIR. This approach was further validated by analyzing pathways of cancer biclusters and prognosis of associated miRNAs followed by confirmatory experiments.

4.3 Materials and Methods

4.3.1 Collection of expression fold-change data

First, the CEL files of 2019 GEO series produced using the Affymetrix U133 Plus 2.0 chip were downloaded¹⁹⁶. Next, Robust Multi-array Average (RMA) normalization was applied to each CEL file using ‘justRMA’ function in R ‘affy’ package¹⁹⁷. The intensities of the probes for each gene were collapsed by their average value. Then, two sample experiments (test/control) were curated for each experimental series and the logarithmic FC (denoted logFC) of the average expression in each group was calculated. In total, logFC profiles for 5,158 (test/control) cell conditions were collected for 20,639 human gene symbols. The logFC matrix and cell condition information is available from our bimir R package (<https://github.com/unistbig/bimir>).

4.3.2 Sequence-specific miRNA targets

The sequence-specific miRNA targets were obtained from the seven sequence-based target prediction databases (TargetScan¹⁹⁸, miRanda¹⁹⁹, mirSVR²⁰⁰, PITA²⁰¹, DIANA-microT²⁰²⁻²⁰³, miRDB²⁰⁴ and TargetRank²⁰⁵). I only used sequence-specific mRNA targets that were reported to have a binding sequence from three or more databases out of the seven. The number of mRNA:miRNA interactions, parameters used, and the download sites for the sequence-specific targets are available from Supplementary information of Chapter IV (‘Collection of sequence-based miRNA targets’ section).

4.3.3 miRNA target prediction using a Progressive Biclustor Extension (PBE) algorithm

The overview of biclustering-based miRNA target prediction is shown in Figure 4.2. First, 5,158 mRNA microarray datasets with two sample groups (test/control) were collected from Gene Expression Omnibus database^{188, 196}, and corresponding logarithmic FC (LFC) data were compiled for 20,639 human genes (columns) and 5,158 FC cell conditions (rows). These LFC data are quantized into up-, neutral-, and down-regulated genes (denoted by 1, 0, and -1, respectively) using $\pm \log_2 1.3$ (FC) thresholds. For each miRNA, sequence-specific targets predicted in at least three out of seven miRNA target databases were selected (denoted as background set).

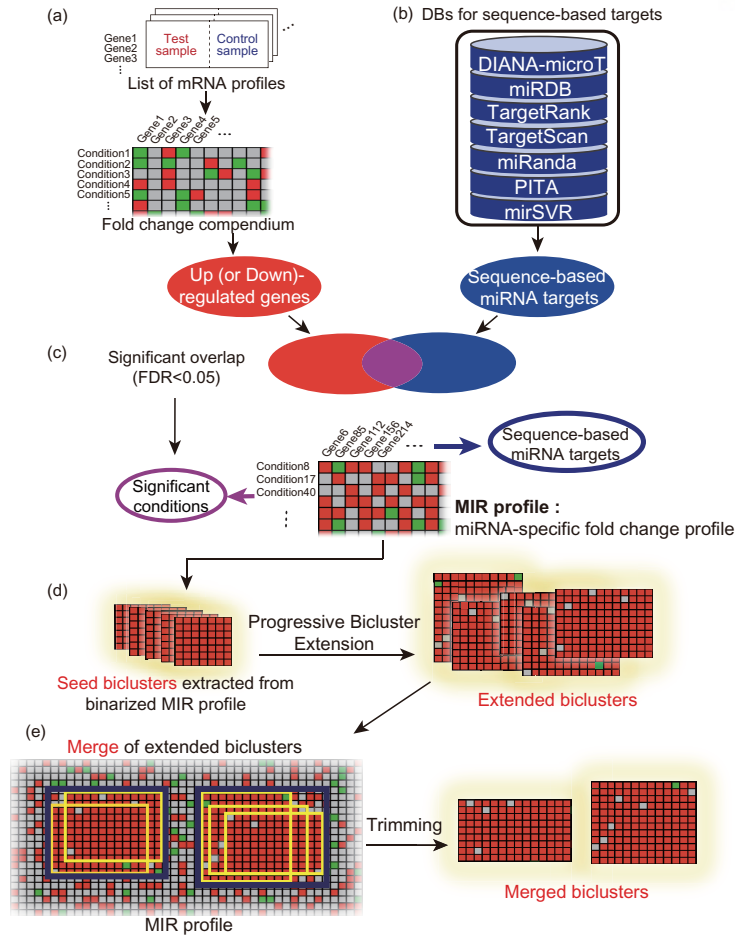


Figure 4.2. Overview of the biclustering-based miRNA target prediction.

(a) The gene expression fold-change compendium. (b) Sequence-specific targets for each miRNA were obtained from seven miRNA target databases. (c) The MIR profile is composed of binarized logarithmic fold-change values of sequence-specific targets for selected cell conditions. (d) From MIR profile, seed biclusters are extracted using BIMAX algorithm, and then are extended using PBE algorithm. (e) Finally, merged biclusters are generated by hierarchical clustering of extended biclusters and removing the noisy rows and columns.

Then, LFC profiles are assigned to the background set based on the enrichment of up-regulated genes in the background set (hypergeometric test, $FDR < 5\%$). The resulting LFC submatrix is converted to a binary matrix by replacing -1 with 0, and is dubbed *MIR profile* for the given miRNA. First, the bimax biclustering algorithm¹⁹¹ is applied to the MIR profile to obtain a number of small biclusters completely filled with 1 (called seed biclusters). These biclusters are then ‘progressively’ extended using PBE algorithm (extended biclusters, see Progressive Bicluster Extension (PBE) algorithm in Supplementary Information of Chapter IV, Figure S4.1); rows and columns with many 1’s are competitively added to the seed bicluster and then relatively noisy rows and columns are removed, and

this process is repeated by slightly increasing the threshold for zero proportion in biclusters (strict to less strict). The extended biclusters are then clustered using average-linkage hierarchical clustering (merged bicluster) to remove redundant results. Testing the three distance cutoffs (0.3, 0.5 and 0.7) for clustering, I found that the cutoff had almost no effect to the result, so the cutoff=0.5 was used. After the merging, the rows or columns that contain more than 10% of zeros are trimmed off one by one to finally yield the 'merged biclusters'. The pseudocode of PBE algorithm is written in Figure S4.2.

The resulting biclusters represent predicted target genes (bicluster columns) up-regulated across a number of cell conditions (bicluster rows). Down-regulated biclusters are also generated in the symmetrical way. Detailed features of the biclusters are described in Table S4.2 and Figure S4.3. Up (down)-regulated biclusters imply the corresponding miRNA is down (up)-regulated in the captured test conditions. The analysis results for $\pm \log 1.3$ thresholds are mainly reported here, but the biclusters were also generated under $\pm \log 1.5$ and $\pm \log 2.0$ thresholds and analyzed. An example of let-7c bicluster for stem cell conditions are described in Supplementary information of this chapter.

4.4 Results

4.4.1 Comparison with existing biclustering algorithms

Compared with seed biclusters, PBE algorithm yields much larger biclusters by allowing for a small portion of noise (Figure S4.3). Its performance was compared with those of five existing biclustering algorithms such as ISA¹⁹², QUBIC²⁰⁶, FABIA¹⁹⁴, BIBIT¹⁹⁵ and HOCCLUS2¹⁸¹ that detect 'up-regulated' constant biclusters. Detailed information of each method is described in the Supplementary information of Chapter IV ('Comparison of biclustering algorithms' section). First, the size and signal density of biclusters generated from a real MIR profile (hsa-let-7c-5p) were compared (Table S4.3). PBE yielded large biclusters with high densities (small proportion of zeros), whereas existing algorithms yielded biclusters with either smaller sizes or poorer densities. PBE also captured the stem cell bicluster better than existing algorithms (Figure S4.4). Detailed result for real data analysis is described in Supplementary Information ('Comparison of biclustering algorithms – Real data analysis' section)

Next, I tested sensitivity and specificity of six biclustering methods using simulation binary profile reflecting the average size and density of real MIR profiles (700 rows, 300 columns and 20% density) (Figure 4.3). The simulation profile contained seven biclusters of which row and column sizes were between 20~80, and each bicluster included 1~3% of zeros (noise). Some of biclusters were overlapped to each other by less than 20% of the bicluster sizes. The simulation was repeated 50 times. Here, 'true' was defined as the elements included in the seven biclusters, and 'false' was the others in the profile. Thus, the sensitivity was defined as the number of true elements within all resulting biclusters divided by the number of all true elements. The precision was defined as the proportion of the true elements within all resulting biclusters.

PBE showed perfect precision (median=100%) with high sensitivity (median=95.6%). The performance of ISA depended on the row (TG) and column (TC) thresholds. When TG=TC=1, it showed high sensitivity (median=97.2%) but relatively low precision (median=87.7%). When both TG and TC increased to 2, the precision was increased (median=96.8%) but the sensitivity was lowered (median=86.1%). The QUBIC results were affected by the consistency parameter c . As this value increased, the precision was increased while the sensitivity was decreased. It showed the best performance with default parameter ($c=0.95$, median precision=80.8%, median sensitivity=100%). BIMAX and BiBit do not allow zeros in the biclusters. When they were run once, they exhibited quite low sensitivity (median BIMAX sensitivity = 10.2%, median BiBit sensitivity = 14.5%). However, the sensitivity of BIMAX increased to 86.7% as it was run 30 times, while that of BiBit was not changed. FABIA yielded very noisy biclusters for all tested sparseness parameters ($a=0.01$ and 0.05) resulting in low precision (median=46.6%) and sensitivity (66.0%). For $a \geq 0.1$, it did not create biclusters. HOCCLUS2 was also tested but excluded in the graph because it didn't generate any bicluster under this simulation setting. These results indicate that the progressive extension process in PBE algorithm is an efficient way to find biclusters from noisy data.

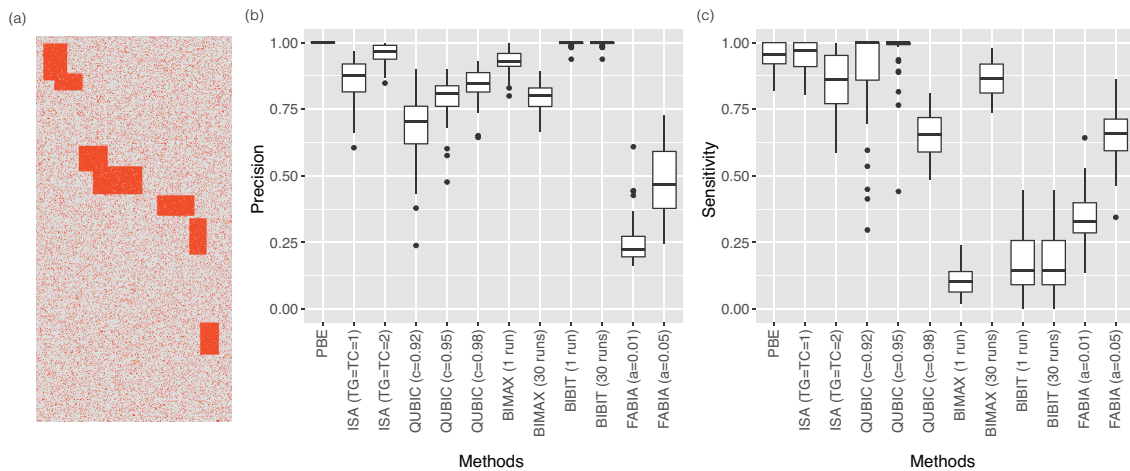


Figure 4.3. Simulation test for biclustering algorithm.

(a) Example simulation profile. Orange and gray elements indicate 1 and 0, respectively. (b) Precision and (c) sensitivity of tested biclustering methods.

4.4.2 Accuracy of the biclustering target prediction

The bicluster targets were assessed using validated miRNA targets. miRTarBase²⁰⁷ provides hundreds of thousands of experimentally validated miRNA-target relations with ‘strong’ evidences (Reporter assays or Western blot) and ‘less strong’ (or weak) evidences (pSILAC or microarray experiment). Among the sequence-specific targets (background set) of a given miRNA, those validated with ‘strong’ evidences were regarded as gold positive (GP) targets, whereas those having neither strong nor weak

evidences were set as gold negative (GN) targets. For evaluation, I selected miRNAs having more than 30 GPs whose fraction within the background set was not less than 5%. 11 miRNAs that satisfied these criteria were analyzed (Figure 4.4a).

For each miRNA, all the resulting bicluster targets, whether up- or down-regulated, were pooled as predicted targets, and corresponding sensitivity, specificity, as well as GP enrichment and GN depletion were calculated (Table S4.5-S4.8). When 1.3 FC threshold was used to quantize the FC data, the average sensitivity and specificity of the 11 miRNAs were 0.704 and 0.466, respectively (summation = 1.170), hence 17.0% (median 19.4%) improved gain compared with the sequence-based target prediction. Although positive gains were obtained for all the 11 miRNAs for 1.3 FC (Figure 4.4a), the relative performances for each miRNA were quite different for different FC cutoffs (Table S4.5). For example, the gain of miR-34a-5p was decreased as the FC cutoff was increased because of the rapid decline in sensitivity (gains for 1.3 FC: 20.8%, 1.5 FC: 13.3%, 2.0 FC: 7.2%). In contrast, the gain of miR-21-5p increased as the cutoff was increased because the specificity was relatively more increased (gains for 1.3 FC: 16.4%, 1.5 FC: 26.5% and 2.0 FC: 31.3%). It presumably represents the different miRNA regulation patterns. The former case corresponds to the ‘fine tuners’ that moderately regulate many genes. Therefore, using lower cutoff helps detect subtle changes in target expressions. However, miRNAs for the latter case seem to more strongly regulate relatively small number of targets. Among the three thresholds, 1.3 FC exhibited the best overall gain with the largest sensitivity.

MiRNA targets tend to be functionally related to each other²⁰⁸. Therefore, I incorporated the protein functional interaction networks from STRING database⁷⁶ (edge threshold ≥ 150) between the bicluster target genes to improve the prediction. Among the bicluster targets, those with k or more functional interactions with other targets were further selected, and the corresponding gains were measured. Intriguingly, the specificity rapidly increased as k was increased (Figure 4.4b), and the maximum gain reached up to 32.0% when $k = 3$ (specificity = 77.8%, Figure 4.4c). The maximum median gain was even higher (33.4% when $k = 4$). These results show that applying the network information considerably improves the miRNA target prediction.

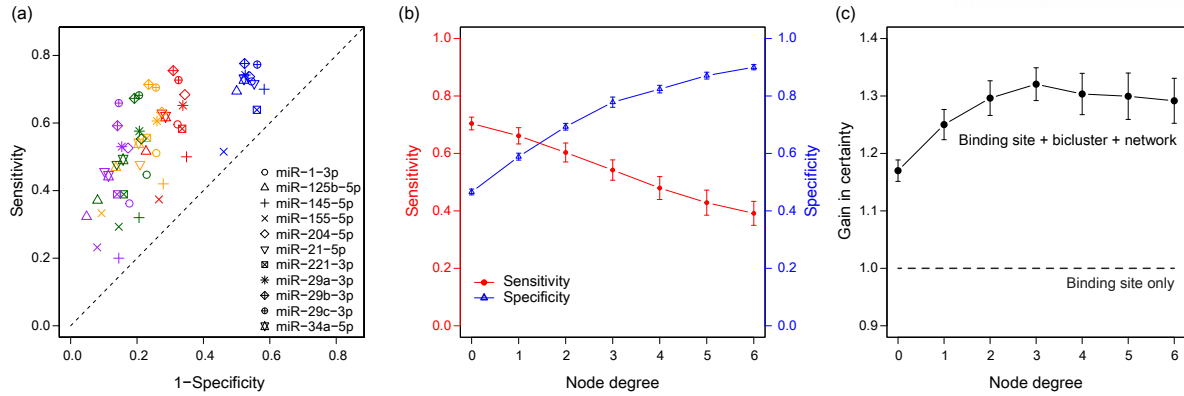


Figure 4.4. Performance of miRNA target prediction using binding sequence, biclustering, and functional networks.

(a) Sensitivity and specificity of pooled bicluster targets of eleven miRNAs. Targets with binding sequence were used as background (diagonal black dash). Blue nodes represent biclustering results. Red/yellow/green/purple nodes represent the results obtained using both the biclustering and network-based target selection with node-degrees 2, 3, 4 and 5, respectively. (b) Average sensitivity and specificity for different node degrees of target networks. (c) Average gains in certainty of methods using binding sequence, biclustering, and network information. Bars in (b) and (c) represent standard error.

4.4.3 Comparison with anticorrelation-based methods in cancer

mRNA-miRNA profiling has been commonly used to predict condition-specific miRNA targets based on the anticorrelation between miRNA and its mRNA targets. Therefore, I compared the biclustering method with seven anticorrelation-based methods (GenMiR++¹⁷⁷, Pearson correlation, Spearman correlation, Lasso²⁰⁹⁻²¹⁰, Elastic Net²¹¹, IDA²¹² and Tiresias²¹³) in predicting cancer-specific miRNA targets. Pearson, Spearman correlation, Lasso, Elastic Net and IDA were implemented using miRLAB R package²¹³⁻²¹⁴ and GenmiR++ and Tiresias were run using MATLAB and Perl software, respectively. For the 11 miRNAs evaluated in the previous section, biclusters where at least 30% of the rows are about ‘tumor vs. normal’ or ‘aggressive vs. non-aggressive tumor’ conditions were selected. These biclusters represented 33 miRNA-cancer pairs for five cancer types (breast, brain, lung, colon, or blood cancer). All of these cancer types had both the mRNA and miRNA data in TCGA, so it was possible to test anticorrelation-based methods. For the biclustering method, I pooled the bicluster targets in the order of proportion of the specific cancer condition in each bicluster. Thus, the true and false positive rates of bicluster targets in each pooling step were shown, while ROC (receiver operating characteristic) curves were depicted for the anticorrelation-based methods (Figure 4.5). After removing six cases where none of the all AUCs (areas under the curves) exceeds 0.6 and the maximum biclustering gain was less than 1.1, twenty cases that were coherent with the known expression of corresponding miRNAs

(quantitative PCR results) were selected for comparison. In other words, up (down)-regulated biclusters were chosen if the corresponding microRNA is known to be down (up)-regulated in cancer. Table S4.9 lists the literature reporting the expression levels of miRNAs in cancers.

Overall, the biclustering method was compared favorably with the mRNA-miRNA profile based methods (Figure 4.5). For 11 out of the 20 cases, the biclustering method exhibited better gains compared with the anticorrelation-based methods; in other 6 cases, both approaches exhibited similar performances; in the remaining 3 cases, the biclustering method was inferior to the best anticorrelation-based method, mostly because of its low sensitivity. As seen in the previous section, incorporating network information tended to increase the specificity (and the gain) of the biclustering method. Among the four anticorrelation-based methods, Genmir++ performed best for most cases.

These results show that our biclustering approach, if miRNA expression information is provided, overall performs better than anticorrelation-based methods in prioritizing condition-specific miRNA targets. The miRNA expression is relatively easily obtained from the literature or quantitative PCR experiment.

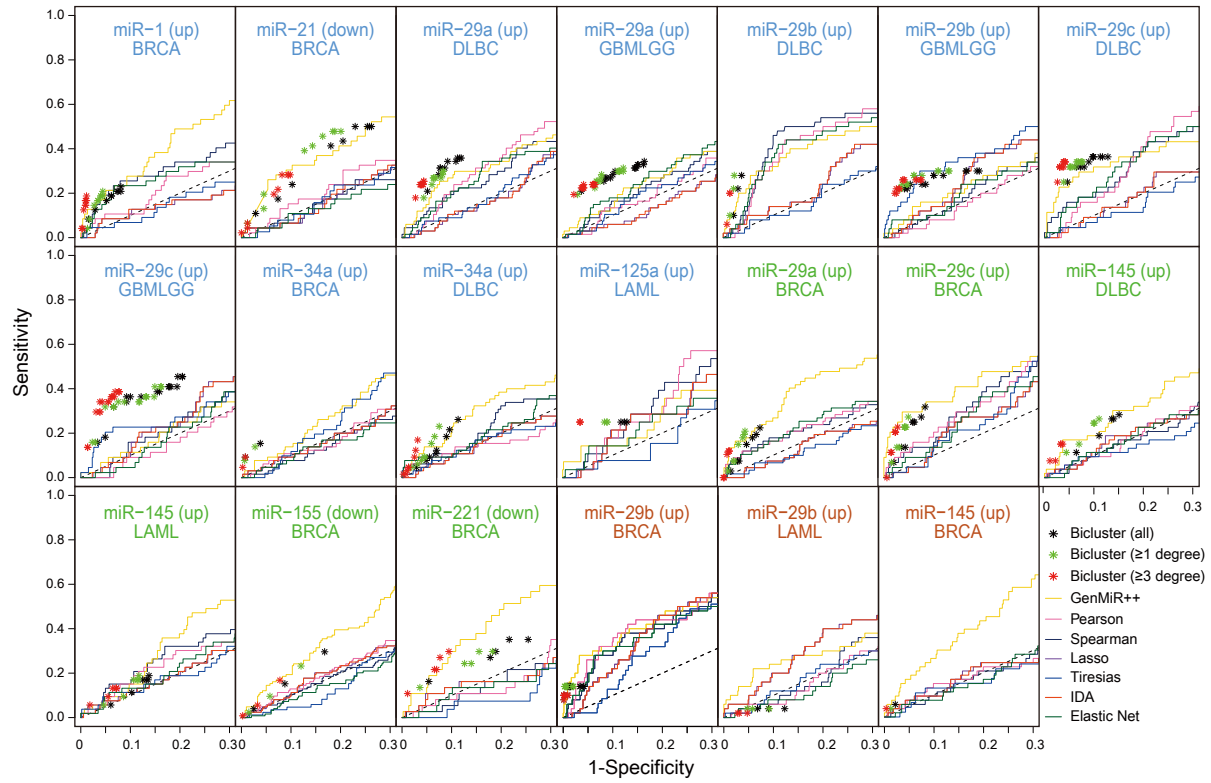


Figure 4.5. Performance comparison between biclustering and anticorrelation-based methods.

Black asterisks represent bicluster predictions. Green and red asterisks represent bicluster targets with at least one and three network degrees, respectively. Solid lines represent ROCs of the seven anticorrelation-based methods. The title of each panel represents the cancer type, miRNA, and target regulation direction (parenthesized). Blue, green, and red titles represent the 11, 6, and 3 cases where the biclustering method performed better than, similar to, and worse than anticorrelation-based methods, respectively. Dashed black lines represent the background results when only sequence-specific targets were used. BRCA, DLBC, GBMLGG and LAML represent breast invasive carcinoma, diffuse large B-cell lymphoma, glioma and acute myeloid lymphoma, respectively.

4.4.4 miRNAs targeting PI3K/Akt signaling in cancer

I further analyzed the bicluster targets corresponding to the 20 cancer-miRNA pairs (Fig. 4.5). Among them, breast cancer and DLBCL yielded the largest numbers of biclusters. In breast cancer, bicluster targets of miR-1, miR-29a/b/c, miR-34a, and miR-145 were upregulated in aggressive cancer; in DLBCL, the targets of miR-29a/b/c, miR-34a, and miR-145 were also upregulated. I pooled those bicluster targets in each cancer type and performed pathway enrichment analysis (KEGG annotation) using the DAVID tool³⁸ to identify six and five significant pathways (FDR<0.05) in breast cancer and DLBCL, respectively (Table S4.10 and S4.11). Interestingly, the bicluster targets in both cancer types were strongly enriched with ‘PI3K/Akt signaling pathway’ (FDR = 1.3E-8 for breast cancer; FDR = 9.1E-8 for DLBCL). This pathway is known to be frequently hyperactivated in many cancers to promote cell cycle and survival, proliferation, and epithelial-mesenchymal transition of tumor cells²¹⁵⁻²¹⁶. In addition, extracellular matrix (ECM)-receptor interaction and focal adhesion pathways were commonly caught in both cancer types, but all the corresponding bicluster targets except two (CAV2, BIRC2) were also included in PI3K/Akt signaling pathway.

Figures 4.6a and S4.5a show PI3K/Akt pathway where the bicluster targets are highlighted for breast cancer and DLBCL, respectively. In both cancer types, the miRNAs targeted multiple ligands including genes encoding growth factors (e.g., VEGFA and PDGFC targeted by miR-29) and ECM (e.g., COL1A1, LAMC1, THBS2 by miR-29); signal transducers such as receptor tyrosine kinase (e.g., MEK and/or PDGFRA by miR-34a), G-proteins (GNB4 and GNG12 by miR-29), toll-like receptor (TLR4 by miR-34a and miR-145) and integrin (e.g., ITGB1 by miR-29); as well as downstream effectors such as NRAS (by miR-29 and miR-145) and CDK6 (by miR-29). In addition, AKT3 was targeted by miR-29 in breast cancer, and cytokine receptor (IL2RB and IL6R) and one component of the PI3K complex (PIK3R3) were also targeted by miR-34a and miR-29, respectively, in DLBCL. Indeed, it was previously shown that miR-29b upregulation in breast cancer considerably inhibited metastasis by repressing targets related to the tumor microenvironment²¹⁷ (including some genes listed above). In the present study, the bicluster targets of miR-29 were experimentally validated using the human breast cancer cell line, MDA-MB 231, which is a well-established metastatic and invasive cancer cell line (done by Woobeen Cho, a Ph.D student in Prof. Ji-young Park’s Lab.). Transcript levels of nine bicluster targets related to ECM or PI3K were analyzed 2 days after transient transfection with either miR-29 or control miRNA. All the nine targets were considerably downregulated by miR-29b or -29c transfection compared to that of the control (Figure 4.6c). Furthermore, the activation of ECM related downstream pathways such as focal adhesion kinase (FAK) and AKT were also considerably attenuated by miR-29 (Figure 4.6d) demonstrating the capability of biclustering analysis to capture relevant pathways for disease. Detailed experimental methods are available in Supplementary Information of Chapter IV.

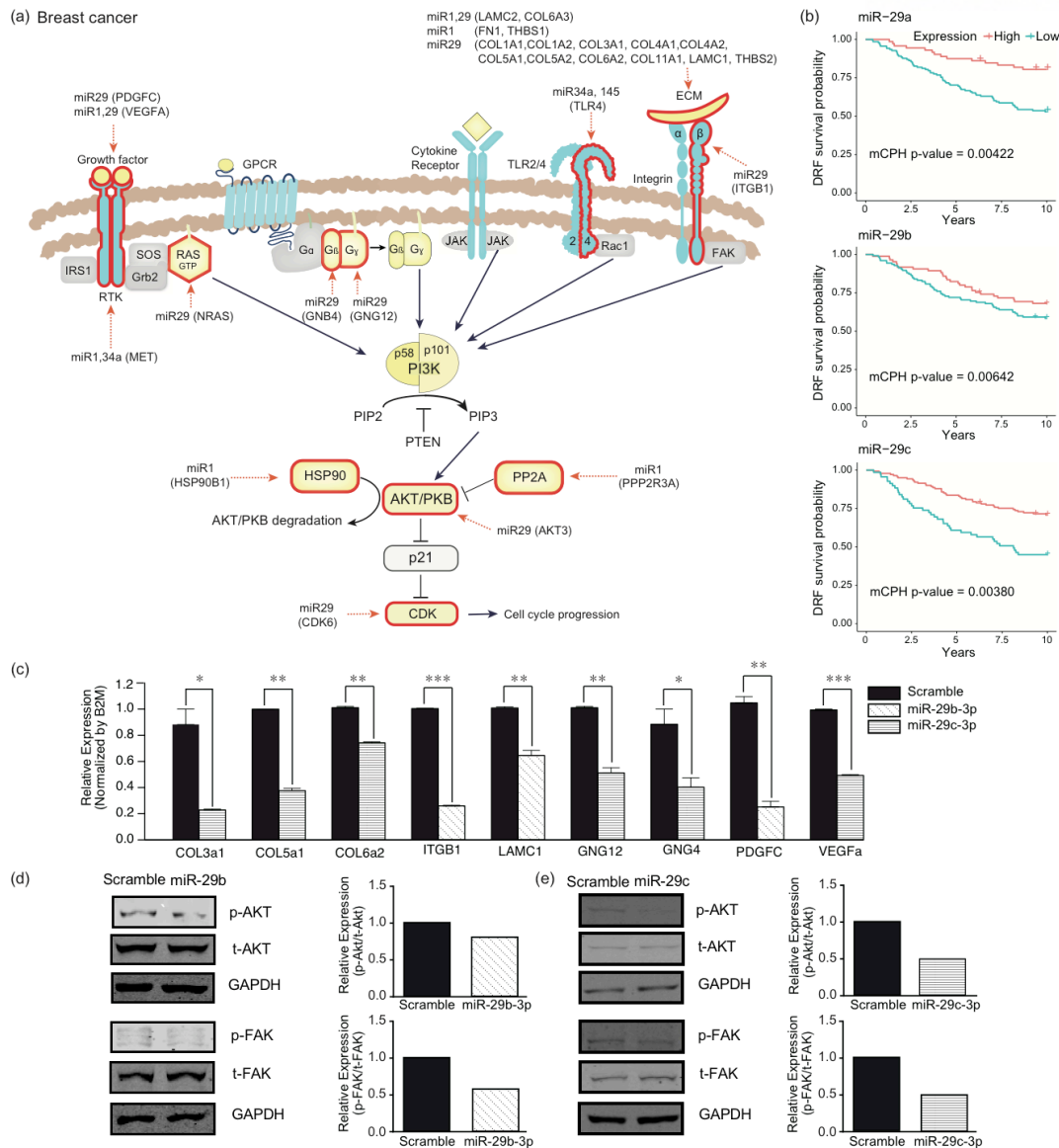


Figure 4.6. miRNA targets in PI3K/Akt pathway (breast cancer).

(a) MiRNA targets predicted from breast cancer biclusters are highlighted by red borders. For each target molecule, corresponding miRNA names and target gene symbols are represented. (b) Distant relapse-free survival analysis for 210 patients with breast cancer exhibiting high and low miR-29a, miR-29b, and miR-29c levels. The patients were divided into two groups based on their best splits at top 33.8%, 40% and 66% values, respectively. (c) Transcript levels of miR-29 target gene candidates were analyzed by qRT-PCR. MDA-MB-231 breast cancer cells were transiently transfected with either scrambled miRNA (control) or miR-29 (29b-3p or 29c-3p). 2-days after transfection, total RNAs were subjected to analyze target gene candidates. Genes were normalized with B2M. All the tested genes were considerably downregulated by miR-29b and/or 29c. ITGB1, LAMC1 and PDGFC were more effectively repressed by miR-29b, and vice versa. (d-e) Activation of downstream pathway candidates such as AKT and FAK were analyzed by immunoblotting. (d) Total cell lysates extracted from either scrambled miRNA or miR-29b-3p (d), as well as miR-29c-3p (e) transfected MDA-MB-231 cells were analyzed for the levels of pAKT, AKT, pFAK and FAK. GAPDH was used as lading control. Quantified results were represented as a bar graph. Statistical significance was evaluated by unpaired one-tailed Student's *t*-test. **p* < 0.05; ***p* < 0.01; ****p* < 0.001 vs. scrambled miRNA.

Finally, I analyzed the prognostic values of these miRNAs using public miRNA expression datasets. The distant-relapse-free survival was tested for 210 patients with breast cancer (GEO database, GSE22216²¹⁸). Among the six miRNAs analyzed, the three miR-29 family miRNAs had significant prognostic values (mCPH *p*-values of miR-29a = 0.0042, miR-29b = 0.0064, miR-29c = 0.0038; adjusted for age, tumor size, lymph nodes involved, ER, and grade). Then, the overall survival of 116 patients with DLBCL (GSE40239¹⁸⁷) was also analyzed for five miRNAs. Among them, two exhibited significant prognostic values (mCPH *p*-values for miR-34a = 0.0185 and miR-145 = 0.0041; adjusted for International Prognostic Index (IPI) and gender). See Table S4.12 and S4.13 for detailed results. Kaplan-Meier plots contrasting the effects of high and low miRNA levels on survival are also shown in Figures 4.6 and S4.5.

Overall, by analyzing cancer biclusters, the key pathways (PI3K/Akt signaling, ECMs, and focal adhesion), and five associated prognostic miRNAs (mir-29a, mir-29b, and mir-29c in breast cancer; mir-34a and mir-145 in DLBCL) that are repressive of tumor progression (hazard ratios 0.593 – 0.745) were identified. In particular, the effects of mir-29b/c on these pathways were experimentally validated.

4.4.5 BiMIR: a bicluster database for condition-specific miRNA targets

In total, 29,815 biclusters were generated for 451 human miRNAs using PBE algorithm (13,921 for 1.3 FC; 10,958 for 1.5 FC, 4,936 for 2.0 FC thresholds) and compiled in BiMIR database (http://www.btool.org/bimir_dir/; constructed by Dr. Hai C. T. Nguyen) where biclusters are searchable for miRNAs, tissues, diseases, keywords, target genes of interest, and their combinations. BiMIR can be used for investigating novel miRNA functions, targets, and related cell conditions.

Along with the list of searched biclusters, the function enrichment results for bicluster targets are provided based on the MSigDB¹¹³ pathway (C2) and gene ontology (C5) categories. If biclusters are searched for a specific organ/tissue or disease, the proportion of corresponding condition in each bicluster is also reported. These help the user to find most relevant biclusters. The heatmaps for each bicluster are visualized (Figure S6) and corresponding target genes and cell conditions are hyperlinked to Genecards^{146, 219} and GEO^{196, 220} databases for detailed information, respectively. For bicluster target genes, the network node degrees, experimental evidences from miRTarBase²⁰⁷, protein network visualization based on STRING database⁷⁶ are provided. In addition, the entire mRNA FC profiles, biclustering R code, and all the biclusters are downloadable from BiMIR database.

4.5 Discussion

Here, a novel framework was presented that prioritize miRNA targets by biclustering sequence-specific targets and cell conditions, a dimension rarely explored before. This is based on the idea that miRNA targets, like other cellular molecules, have a modular activity and will be repeatedly captured across different cell conditions. Indeed, the bicluster targets exhibited substantially improved accuracy compared with purely sequence-based targets and were often enriched with well-known pathways characterizing the modules identified. Moreover, the functionally connected targets exhibited even higher accuracy, further confirming the modular activity of miRNA targets.

I analyzed cancer biclusters and found that PI3K/Akt signaling pathway was intensively targeted by a few miRNAs in two cancer types. Further, prognostic values of those miRNAs and the regulatory effects of mir-29 were also validated. These results demonstrate that biclustering analysis is able to reveal key pathways regulated by miRNAs in disease. BiMIR database provides miRNAs and targeted pathways for dozens of diseases.

Given the miRNA expression, the prediction using biclustering method was favourably compared with seven anticorrelation-based methods in cancer conditions. This demonstrates the practical value of this approach in that bicluster results can provide fairly good target predictions for a variety of cell conditions without generating costly paired expression profiles. BiMIR database was designed so as to explore the modular regulatory networks of miRNAs by connecting miRNAs, cell conditions (or disease), mRNA targets, and associated pathways. The user may find candidate miRNA and target genes for the cell condition of interest. The knowledge of miRNA expression level will help select the right direction of biclusters (up or down).

Despite the improvements and usefulness shown in this study, there remain difficulties in our approach regarding free parameters that need to be optimized. First, the minimum seed size of 10 by 10 was determined in an *ad hoc* manner, and its optimal size may be affected by the size of the fold-change data. Second, the iteration number of 20 in BIMAX algorithm was used to compromise the computation time; using a higher iteration number yielded more biclusters. However, other parameters seemed to be less sensitive. For example, I slightly increased the threshold of zero proportion from 0.01 to 0.1 (step size 0.01) during ten iterations of bicluster extension. This may seem to allow 10% of zeros in the end, but the final zero proportion was only about 1.5% because of the trimming process. The cutoff of hierarchical clustering of the extended clusters was also a less sensitive parameter. In addition, the biclusters were generated under a rather strict criterion (for targets in three or more databases); therefore, BiMIR can be used for selecting a small number of highly likely targets for the cell condition of interest. The biclustering approach presented here can also be applied for predicting the condition-specific targets of other sequence-specific regulators such as transcription factors or RNA binding proteins. In this regard, the entire 5,158 mRNA fold-change profiles for 20,639 genes are provided for general

systems biology research. These mRNA fold-change data are different from the GTEx transcriptome data²²¹ in that GTEx data represent transcription levels in normal tissues, whereas our fold-change data represent gene expression ‘changes’ for a variety of cell conditions such as disease, chemical treatment, tissues, and differentiations. Thus, these fold-change data can also be used for clustering or regulatory network analysis for a specific group of genes or cell conditions.

Another possible future work is coregulatory network of miRNAs. Whereas existing methods to identify miRNA regulation modules bicluster multiple miRNAs and multiple target genes representing coregulatory networks, my current work is focused on prioritizing highly likely target genes of a single miRNA commonly detected across multiple cell conditions. This approach can also be extended to tackle the miRNA coregulatory networks by overlapping biclusters for different miRNAs. A significant overlap implies coregulated mRNA targets under multiple cell conditions. I hope that this approach and data contribute to disentangling the modular structure of complex regulatory networks.

4.6 Supplementary information of Chapter IV

Table S4.1. Existing miRNA target prediction tools

Sequence-based target prediction methods		
Method	Features used in target prediction	References
TargetScan	Seed match, Conservation	198
PITA	Seed match, Conservation, Free energy, Site accessibility, Target-site abundance	201
miRDB	Seed match, Conservation, Free energy, Machine learning	204
mirSVR	Seed match, Conservation, Free energy, Site accessibility, Machine learning	200
miRanda	Seed match, Conservation, Free energy	199
DIANA-microT-CDS	Seed match, Conservation, Free energy, site accessibility, Target-site abundance, Machine learning	202-203
TargetRank	Seed match, Conservation, Base composition at position t9, flanking AU content	205
Correlation/Causality-based target prediction methods		
Method	Features	References
Pearson correlation	Pearson correlation between an mRNA and miRNA	222
Spearman correlation	Spearman correlation between an mRNA and miRNA	223
Lasso	Lasso regression coefficient between an mRNA and miRNA	209-210
ElasticNet	ElasticNet regression coefficient between an mRNA and miRNA	211
GenMIR++	Bayesian learning algorithm	177
Tiresias	Two-stage artificial neural network	213
IDA	Causal structure learning and causal inference	212
Biclustering-based target prediction methods		
Method	Features	References
BIMIR	Biclustering sequence-specific targets and cell conditions using large log expression fold change table.	-
HOCCLUS2	Biclustering mRNA and miRNA using mRNA:miRNA interaction score matrix	181
miRmap	Biclustering mRNA and miRNA using mRNA:miRNA correlation matrix	224
cMonkey2	Biclustering gene expression table and miRNA binding site enrichment test for bicluster genes	225

Data collection

- 1) *Collection of expression fold-change data:* Described in the Materials and Methods section in Chapter IV.
- 2) *Collection of sequence-based miRNA targets*

The sequence-based miRNA targets were set as those predicted from three or more miRNA target prediction databases listed below.

- TargetScan (version 7.0): TargetScan data (Conserved site context++ scores) provided 253,132 miRNA-target interaction data. It was downloaded from TargetScan homepage (<http://www.targetscan.org>).
- PITA (version 6): PITA data (PITA_targets_hg18_0_0_ALL.txt) provided 4,095,751 miRNA-target interaction data. Among them, 716,486 interactions were used of which free energy scores were less than -10. The data was downloaded from https://genie.weizmann.ac.il/pubs/mir07/mir07_data.html.
- miRDB (version 5.0): miRDB data (miRDB_v5.0_prediction_result.txt) provides 1,873,265 miRNA-mRNA interaction data. Among them, 1,314,352 interactions were used of which scores were greater than 60. The data was downloaded from <http://www.mirdb.org/download.html>.
- mirSVR: mirSVR Targets provided 728,288 miRNA-target interactions. The data was downloaded from <http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/> but not available now.
- miRanda (microRNA.org) : miRanda provided 1,097,064 conserved miRNA-mRNA interactions with high mirSVR score (human_prediction_S_C_aug2010.txt). The data was downloaded from <http://www.microrna.org/microrna/getDownloads.do> but not available now.
- DIANA-microT-CDS (version 5.0): DIANA-microT-CDS provides 7,337,705 miRNA-mRNA interactions. Among them, 1,457,011 interactions were used of which scores were larger than 0.7.
- TargetRank: TargetRank provides 1,006,494 miRNA-mRNA interactions. The data was downloaded from (http://hollywood.mit.edu/targetrank/hsa_miRBase_miR_ranked_targets.txt).

Progressive Biclustor Extension (PBE) algorithm

The overall process of PBE algorithm is shown in Figure S4.1 (graphical scheme), and Figure S4.2 (pseudocode). PBE algorithm is composed of two parts: the extension step and the trimming step. Briefly, the seed bicluster is extended by adding the background rows or columns that have the minimum zero rate (extension step) and then noisy rows and columns (showing high zero rate) of the extended bicluster are removed (trimming step). This two-step process is applied R times, and the bicluster is also updated R times accordingly (in this study, $R=10$). The final zero rate allowed in the extended bicluster (Z_{cut}) is set as 10%, but note that the final zero rate was only less than 1.5% on average (Fig S4.3).

Extension step. In the s^{th} step of extension ($s = [1, \dots, R]$), the intermediate zero rate ($Z_{\text{cut},s}$) allowed in extending the current bicluster is defined as:

$$Z_{\text{cut},s} = \frac{Z_{\text{cut}}}{R} \times s$$

For example, if $Z_{\text{cut}}=0.1$ and a seed bicluster is extended through $R=10$ steps, the $Z_{\text{cut},s}$ for the first extension step will be $0.1 * 1/10 = 0.01$. In other words, stricter criteria are applied in the earlier extension steps to obtain biclusters with high densities. Let M be the matrix of the MIR profiles, and $R(s)$ and $C(s)$ be the indexes of the rows and columns of the bicluster that s^{th} extension step is done, respectively. $M[R(0), C(0)]$ denotes the seed bicluster. After calculating the zero rates in every column vector in $M[R(s-1), C(s-1)^c]$ and row vector in $M[R(s-1)^c, C(s-1)]$, the rows or columns with the minimum zero rate are added to the current bicluster. The same extension process is repeated until the zero rate reaches $Z_{\text{cut},s}$, when the bicluster enters the trimming step.

Trimming step. If any row or column vector with the maximum zero rate exceeds $Z_{\text{cut},s}$, such vector is removed from the bicluster one by one resulting in the updated bicluster $M[R(s), C(s)]$.

Prevention of lengthening out in one direction. Some biclusters tend to keep lengthening out in one direction if one side of the bicluster becomes too small compared with the other side during the extension process. To ameliorate this, a penalty is given to the longer side if it is more than twice longer than the other side. When the row and column vectors outside the bicluster compete with each other, the following modified zero rate is applied for the longer side vectors.

$$\text{modified zero rate} = \frac{\# \text{ zeros in the vector} + \text{floor}(r)}{\text{length of the vector}}$$

where,

$$r = \frac{\text{length of longer side of bicluster}}{\text{length of shorter side of bicluster}}$$

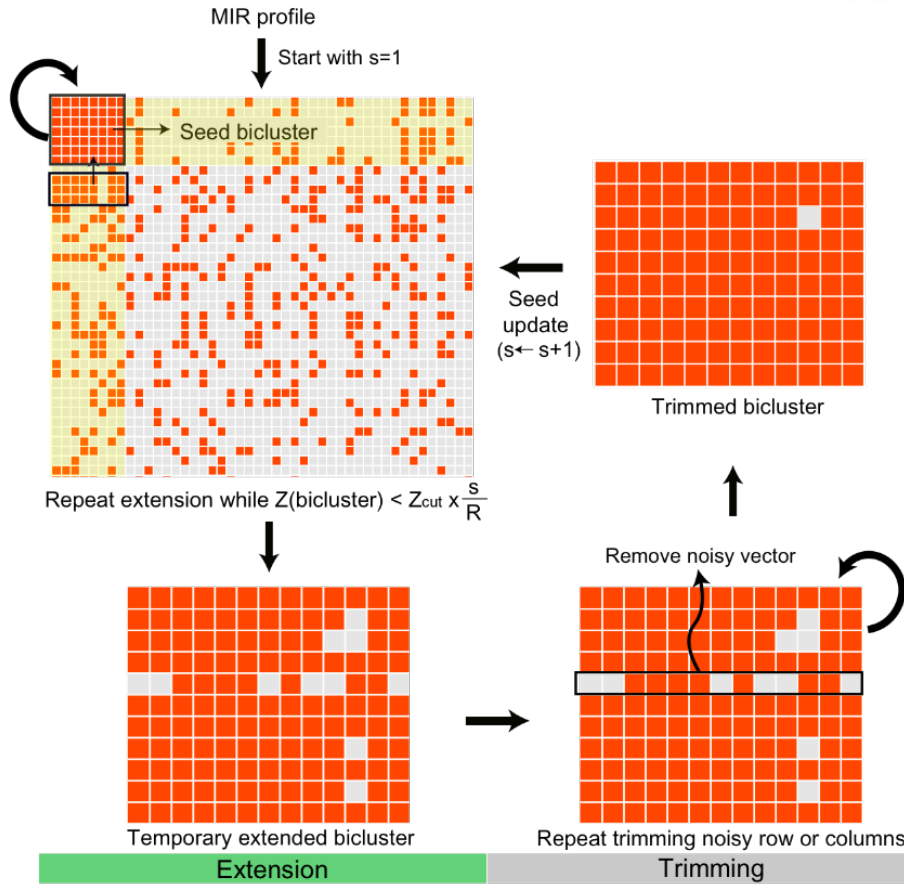


Figure S4.1. Progressive bicluster extension (PBE) algorithm.

The final zero rate cut off (Z_{cut}) in the extended bicluster should be determined in advance. In the MIR profile, the orange and the grey cells represent 1 and 0 respectively, and the seed bicluster is represented in the black box. The seed bicluster is extended by repeating the following extension and trimming process R times. (Extension step) Among the addable vectors (in the yellow shadow), those with the minimum zero rate are simultaneously attached to the current bicluster. If the zero rate in the extended bicluster is less than intermediate zero rate cut-off ($Z_{\text{cut},s} = Z_{\text{cut}} * s/R$, s means s^{th} repetition step), the extension process is repeated. (Trimming step) If the zero rate exceeds the $Z_{\text{cut},s}$, the rows and columns whose zero rate is larger than $Z_{\text{cut},s}$ are searched for and removed from the most noisy vectors to yield the updated bicluster. The updated bicluster enters next extension/trimming step with updated parameter.

Figure S4.2. Pseudocode of Progressive Bicluster Extension

Algorithm : Progressive Bicluster Extension

Input: V_r (condition), V_c (target gene); M^{lsm} (MIR profile); parameters S (The number of extension steps), Z (Final zero rate allowed in the bicluster)

for ($i=1$ to S) **do**

$$Z_{temp} = (Z/S) \times i$$

SEED $\leftarrow M[V_r][V_c]$

Seed extension process

function zero_ratio(array A)

return (The # of zeros in A) / $|A|$

end function

function modified_zero_ratio(array A, integer N)

return (The # of zeros in $A + N$) / $|A|$

end function

while (zero ratio of **SEED** < Z_{temp}) **do**

$V_r \leftarrow$ Conditions in seed bicluster

$V_c \leftarrow$ Target genes in seed bicluster

RowCandidates $\leftarrow M[(V_r)^c][V_c]$

ColumnCandidates $\leftarrow M[V_r][(V_c)^c]$

$n_1 \leftarrow |V_r|/|V_c|$; $n_2 \leftarrow 1/n_1$

if ($n_1 \geq 2$) **then**

$Row_zero \leftarrow$ Values from **modified_zero_ratio()** for all row vectors in **RowCandidates** with $N=n_1$

else

$Row_zero \leftarrow$ Values from **zero_ratio()** for all row vectors in **RowCandidates**

end if

If ($n_2 \geq 2$) **then**

$Col_zero \leftarrow$ Values from **modified_zero_ratio()** for all column vectors in **ColumnCandidates** with $N=n_2$

else

$Col_zero \leftarrow$ Values from **zero_ratio()** for all column vectors in **ColumnCandidates**

```

end if

Min_row_zero ← minimum of Row_zero

Min_col_zero ← minimum of Col_zero

New_conditions ← conditions (rows) in M that corresponds to Row_zero==Min_row_zero

New_target_genes ← Targets (columns) in M that corresponds to Col_zero==Min_col_zero

L1 = |New_conditions|; L2 = |New_target_genes|

if (Min_row_zero < Min_col_zero) OR (Min_row_zero==Min_col_zero AND L1>=L2) then

    SEEDtemp ← M[Vr ∪ New_condition] / [Vc]

else

    SEEDtemp ← M[Vr] / [Vc ∪ New_target_genes]

end if

if (zero ratio of SEEDtemp > Ztemp) then

    break

else

    SEED ← SEEDtemp

    Vr ← Vr ∪ New_conditions

    Vc ← Vc ∪ New_target_genes

end if

end while

# Bicluster Trimming Process

Row_zero ← Values from zero_ratio() for all row vectors in SEED

Col_zero ← Values from zero_ratio() for all column vectors in SEED

Max_row_zero ← maximum of Row_zero

Max_col_zero ← maximum of Col_zero

While (max_row_zero > Ztemp OR max_col_zero > Ztemp) do

    if (max_row_zero >= max_col_zero) then

        conditions_to_delete = SEED conditions (rows) whose zero ratios are equal to max_row_zero

        Vr ← Vr - Conditions_to_delete

        SEED ← SEED[Vr] / [Vc]

```

```

else

    targets_to_delete = SEED target genes (columns) whose zero ratios are equal to max_col_zero

     $V_c \leftarrow V_c - \text{targets\_to\_delete}$ 

     $SEED \leftarrow SEED[V_r][V_c]$ 

end if

    Row_zero  $\leftarrow$  Values from zero_ratio() for all row vectors in SEED

    Col_zero  $\leftarrow$  Values from zero_ratio() for all column vectors in SEED

    Max_row_zero  $\leftarrow$  maximum of Row_zero

    Max_col_zero  $\leftarrow$  maximum of Col_zero

    If(max_row_zero <  $\mathbf{Z}_{temp}$  AND max_col_zero <  $\mathbf{Z}_{temp}$ ) then

        break

    end if

end while

end for

Return SEED

```

Bicluster statistics

By progressively extending the seed biclusters and merging similar ones, many of missing associations can be restored to yield biologically meaningful results. Figure S4.3 represents the distributions of bicluster size (number of conditions and genes) and density (1-zero ratio; 1.3-fold cut-off). For 1.3-fold cutoff bicluster, 11.5 conditions and 10.5 genes were included in the seed biclusters on average. After extending them, the average number of conditions and genes were increased to 19.4 and 28.4, respectively. Finally, merged biclusters had slightly more increased sizes. However, the zero ratio of the merged biclusters was only less than 1.5% on average. Increasing FC cutoff resulted in less extended but slightly denser biclusters (Fig S4.3). Compared with other biclustering methods, PBE was able to identify larger and/or cleaner biclusters from noisy data as shown in the next section.

BiMIR (http://btool.org/bimir_dir/) provides 29,898 biclusters for 459 human microRNAs. These biclusters cover in total 2,259 fold change (FC) conditions (~43% of total cell conditions). Table S4.2 shows six statistics of BiMIR biclusters for three binarization cutoffs (1.3, 1.5 and 2.0 FC). Note that for each miRNA, six MIR profiles were generated (up- and down-regulated profiles for three FC cutoffs). If no biclusters were generated from MIR profile, corresponding miRNA was not counted in Table S4.2.

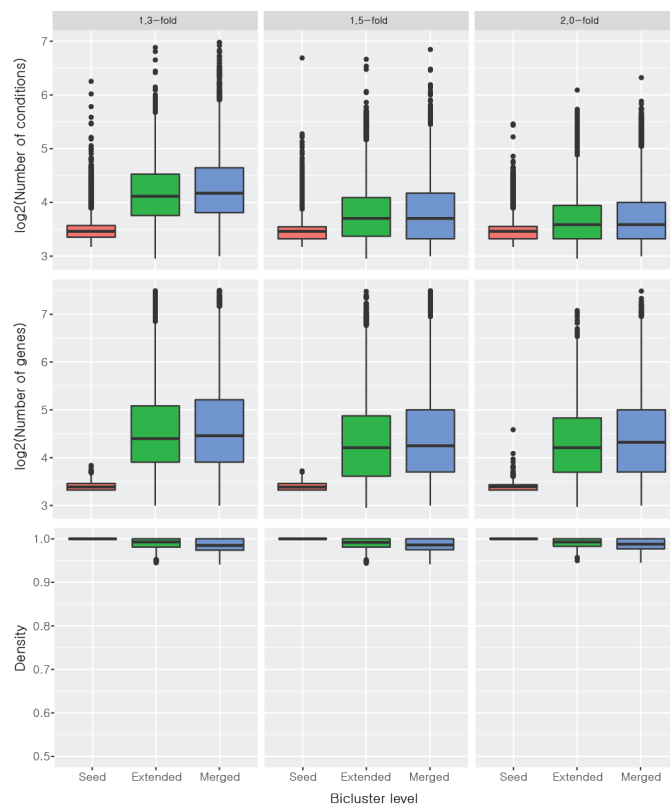


Figure S4.3. Distribution of the number of conditions, genes and density in biclusters with three different fold-change cut-offs.

Table S4.2. Statistics of BiMIR biclusters.

Binarization cut-off	1.3 FC	1.5 FC	2.0 FC
Number of miRNAs	459	414	348
Number of biclusters	13,949	10,999	4,950
Number of FC conditions	2,259	1,828	1,057
Average number of conditions	20.5	15.3	14.1
Average number of genes	30.8	26.7	26.8
Average bicluster density	0.985	0.986	0.988

Comparison of biclustering algorithms

The performance of PBE and other ‘up-regulated’ constant biclustering algorithms were compared in two ways: (1) the size and density comparison using MIR profile of hsa-let-7c-5p and (2) precision and sensitivity comparison using simulated data. In this section, I describe the tested biclustering algorithms, and real data analysis result. Simulation analysis is described in Chapter IV (refer to 4.4.1 Comparison with existing biclustering algorithms).

(1) Compared biclustering algorithms

- Iterative signature algorithm (ISA)¹⁹²: It was developed to find transcriptional modules from microarray gene expression profiles. It aims to detect a set of genes showing similar up- or down-regulation patterns across a set of samples. To achieve the modules, ISA iteratively updates the rows (genes) and columns (conditions) that satisfies the criterion until the result converges. ISA has two parameters: row (T_G) and column (T_C) threshold parameters. In this study, both parameters were adjusted from 1 to 3 by 0.5. It was run by ‘isa’ function in ‘isa2’ R package
- QUBIC²⁰⁶: It is a qualitative or semi-quantitative biclustering algorithm. It automatically converts the continuous input gene expression matrix into signed integer matrix based on the parameters r (e.g., 1=up-regulated, 0=not regulated and -1=down-regulated) and then constructs the gene network in which the edges represent the number of co-regulated conditions. It finds non-overlapping seed biclusters from this network and expand the biclusters based on the consistency parameter that controls the ratio of identical non-zero values in each column. In this study QUBIC biclusters were generated using BCQUD function in QUBIC R package with three consistency levels ($c=0.92$, 0.95 and 0.98).

- FABIA¹⁹⁴: It is a generative multiplicative model designed for gene expression data considering the heavy tails in the distribution. It returns biclusters with ranks evaluated according to the information content. FABIA biclusters were generated using ‘fabia’ function in fabia R package with sparseness loading parameters 0.01, 0.05, 0.1, 0.15, 0.2, 0.25 and 0.3. Parameter p (the number of bicluster) was set as 30 for real data analysis and 7 for simulation data.
- BiBit¹⁹⁵: it was developed for biclustering of binary matrix. It transforms input binary matrix to integer matrix by dividing every rows into bit words of same size and then converting each bit word into decimal number. It is fast by searching biclusters from this reduced integer matrix. It was run using ParBiBit program²²⁶ which accelerated the running time of BiBit algorithm by implementing MPI parallel programming.
- HOCCLUS2¹⁸¹: it was developed to bicluster microRNAs and target genes on binary data (experimentally validated or predicted interaction networks). In the first step, the initial bi-cliques are generated based on the minimum interaction score. Then, the overlapping biclusters are progressively merged based on the cohesiveness parameter which measures the quality of each bicluster by the functional similarity of genes. HOCCLUS2 has two input parameters such as α (a cohesiveness threshold) and β (a minimum interaction score). Because $0 < \beta < 1$ does not affect the result when applied to a binary data, it was fixed to 0.5 and only α was changed from 0.4 to 0.9.

(2) Real data analysis

The up-regulated MIR profile of hsa-let-7c-5p (FC cut-off=log2(1.3); 1526 conditions x 801 genes) was used to compare the performance of different algorithms. Table S4.3 shows the average row/column size, density and the number of detected biclusters of each algorithm. PBE identified 17 biclusters having average sizes of 23.3 conditions and 38.9 genes and 98.1% density. ISA was applied to both continuous FC data and binarized data (based on 1.3-fold). The density of biclusters from continuous data were estimated using 1.3-fold cut-off. ISA generated smaller and denser biclusters as the T_C and T_G were increased. For example, when both parameters were set as 1, the average numbers of conditions and genes were as large as 174 and 119, respectively, but the density was quite low (50.0%) for biclusters from continuous profile. When both parameters were set as 3, the average density increased to 80.8%, but the average size was quite small (26.1 conditions and 9.2 gene). The average size of BiMIR bicluster was between those of ISA (continuous) biclusters with parameters $T_C=2.5$, $T_G=1.5\sim 2.0$. In that case, however, the average density was much lower than that of PBE (70.3%~71.9%). When same parameters were applied to binary data, it usually generated larger but sparser biclusters than those from continuous profile. QUBIC was implemented with three consistency parameter c (c=0.98, 0.95 and 0.92). When c=0.92, the QUBIC biclusters included the largest number of 1 on

average with quite high density (row size=15.2, column size=61.3, density=97.96%). It tended to contain lesser rows and more columns compared to PBE. It is because to extend rows satisfying the consistency level (the minimum ratio of 1 in each column) for all columns in the seed bicluster is not that easy. It seems appropriate for finding relatively small number of genes co-regulated under large number of samples, but not for our case to find biclusters including many target genes as well as many conditions. Also, QUBIC biclusters sometimes contain noisy rows (See the undermost row in fig. S4.4b). FABIA was tested for both continuous and binarized profiles using various sparseness parameters ($\alpha=0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$). The continuous biclusters were quite sparse for all conditions (density=31.8%~38.8%). The binarized biclusters showed higher density compared to that of continuous biclusters (35.8%~72.3%), but it was still sparser than that of PBE. BiBit resulted in huge number of small biclusters full of 1 (1227 biclusters, row size=13.5 and column size=12.4).

For HOCCLUS2 biclusters, the size and density depended on the bicluster overlapping levels and α (cohesiveness). For all α , the level 1 biclusters were small (row size=13, column size=10) with 100% density. Except for the cases of $\alpha = 0.8$ and 0.9 with which biclusters were barely extended, most biclusters became larger containing more zero proportions as the level was increased. The level 2 biclusters had 23 conditions and 19 genes on average for $\alpha = 0.4 - 0.7$ and were most similar to those of PBE biclusters. However their average density was rather lower (84.4% ~ 86.3%) compared with that of PBE biclusters (98.1%). From level 3 ($\alpha = 0.4 - 0.7$), the biclusters showed very low densities which were far from useful to predict regulatory modules (Table S4.3).

All methods found the homogenous biclusters that mostly consist of ESC/iPSC vs. somatic cell conditions. PBE showed the best performance with respect to size and/or density (51 conditions and 126 genes with 97.6% density) compared with other methods. The largest ISA bicluster had 71 conditions and 154 genes with only 83.4% density when $T_C = 2, T_G = 1$, and the densest one had only 35 conditions and 37 genes with 95.6% density when $T_C = 4, T_G = 2$. QUBIC bicluster had 47 conditions and 109 genes with 98.2% density. FABIA generated big bicluster (44 conditions and 293 genes) but the density was very low (80.2%) and BiBit yielded small bicluster (23 condition and 35 genes) with full of 1. The level 2 bicluster of HOCCLUS2 had only 26 conditions and 84 genes with 97.3% density (tests with $\alpha = 0.4 \sim 0.9$ yielded same results) (Fig S4.4).

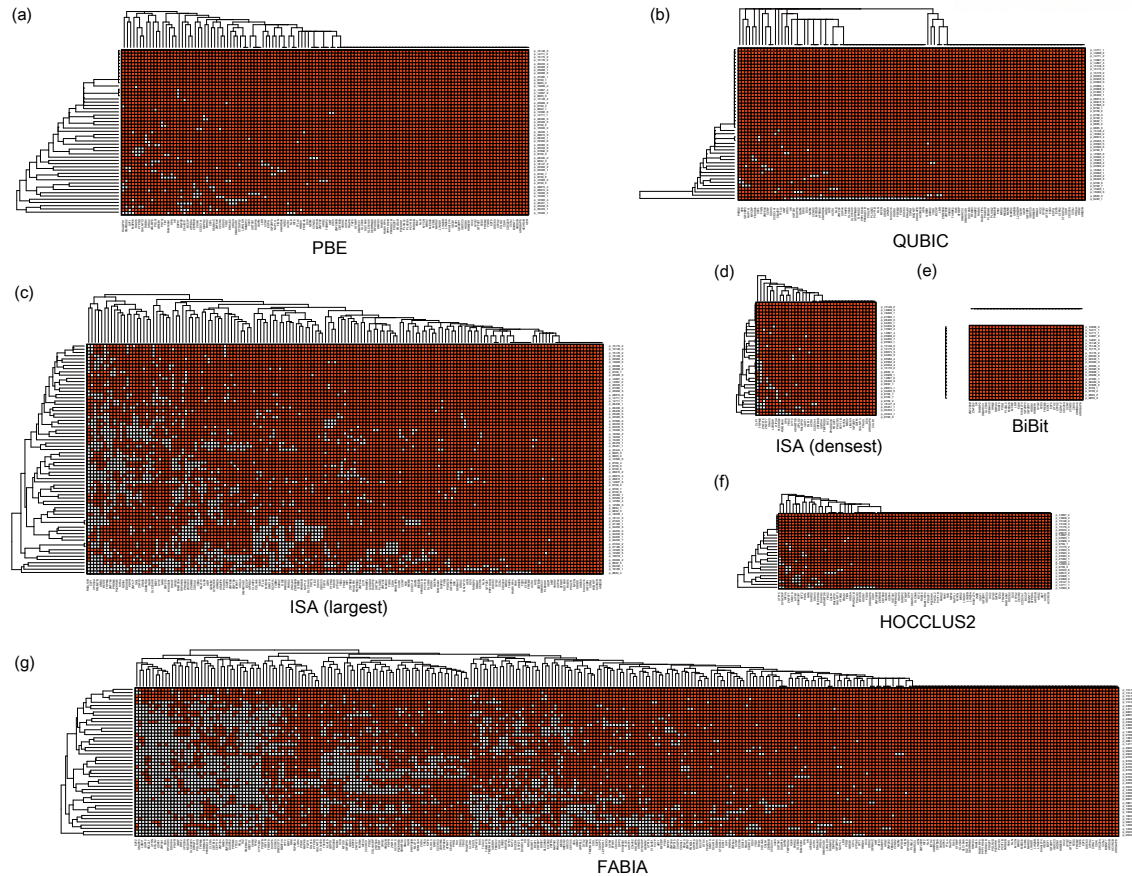


Figure S4.4. ESC/iPSC biclusters searched by multiple biclustering methods.

All biclustering methods detected biclusters containing homogeneous ESC/iPSC vs. somatic cell condition. (a) PBE detected large and dense bicluster (51 conditions and 126 targets with 97.6% density). (b) QUBIC detected rather small but dense bicluster (47 conditions and 109 targets with 98.2% density) (c) ISA found large but noisy bicluster (71 conditions and 154 targets with 83.4% density). (d) The densest ISA biclusters showed relatively small size (35 conditions and 37 targets with 95.6% density). (e) BiBit detected a small bicluster full of 1 (23 conditions and 35 genes with 100% density) (f) HOCCLUS2 found dense bicluster but the size was quite small (26 conditions and 84 genes with 97.3% density) (g) FABIA detected large but very noisy bicluster (44 conditions and 293 targets 80.2% density)

Table S4.3. Real data analysis.

For PBE, QUBIC, BiBit, FABIA, ISA and HOCCLUS2, the average size of row/column, density, and the number of biclusters were measured using up-regulated MIR profile of hsa-let-7c-5p.

PBE									
		Row		Column		Density		N	
		23.3		38.9		98.1		17	
QUBIC									
Consistency		Row		Column		Density		N	
1.0		17.2		41.9		1.0		46	
0.98		16.9		42.0		0.9997		44	
0.95		14.2		43.4		0.9989		36	
0.92		15.2		61.3		0.9796		24	
BiBit (minimum row and column size = 10)									
		Row		Column		Density		N	
		13.5		12.4		1.0		1227	
FABIA									
Continuous input					Binary input				
Sparseness loading		Row	Column	Density	N	Row	Column	Density	N
0.01		80.1	219.4	0.318	30	55.5	226.3	0.358	22
0.05		64.1	195.5	0.330	26	25.8	291.1	0.548	27
0.1		25.9	189.6	0.355	26	15	285.2	0.653	25
0.15		28.5	198.3	0.359	28	10.9	265.6	0.714	23
0.2		25.8	198.7	0.368	25	12.2	268.1	0.723	15
0.25		19.1	201.2	0.388	27	14.4	272	0.701	11
0.3		28.2	195.8	0.361	23	11.9	266.8	0.719	10
ISA									
Continuous input					Binary input				
T _G	T _C	Row	Column	Density	N	Row	Column	Density	N
1.0	1.0	174.0	119.3	0.500	4	192.7	116.2	0.464	6
	1.5	176.7	60.7	0.526	7	191.8	70.2	0.498	9
	2.0	196.5	27.7	0.493	15	200.8	39.5	0.534	13
	2.5	202.8	11.4	0.546	22	216.9	18.9	0.532	22
	3.0	189.5	5.6	0.672	28	221.0	10.2	0.582	18
1.5	1.0	106.4	118.6	0.486	7	95.0	118.0	0.486	9
	1.5	106.7	60.1	0.530	12	94.8	66.8	0.511	9
	2.0	100.1	28.3	0.582	15	113.2	39.5	0.560	11
	2.5	101.6	12.4	0.623	22	127.7	21.2	0.591	24
	3.0	105.2	6.2	0.707	23	133.6	11.1	0.647	16
2.0	1.0	58.1	112.0	0.482	11	59.2	113.8	0.509	12
	1.5	52.5	58.4	0.554	17	58.3	69.5	0.533	13
	2.0	52.3	27.0	0.621	21	72.4	43.2	0.605	11
	2.5	54.3	12.3	0.641	28	66.6	20.9	0.569	29
	3.0	59.2	8.1	0.744	18	74.1	13.0	0.676	18
2.5	1.0	25.8	110.8	0.529	30	28.6	109.0	0.529	24
	1.5	25.8	58.3	0.632	28	32.7	71.3	0.581	23
	2.0	32.0	28.3	0.703	22	33.0	42.1	0.599	29

	2.5	34.9	14.3	0.719	21	38.6	22.0	0.635	30
	3.0	37.2	8.7	0.770	13	45.7	10.1	0.694	44
3.0	1.0	14.7	105.2	0.638	37	15.0	120.6	0.619	41
	1.5	17.1	52.7	0.701	33	16.1	72.5	0.645	45
	2.0	18.3	27.5	0.744	32	18.4	42.7	0.658	42
	2.5	23.0	14.9	0.804	19	21.2	25.2	0.654	46
	3.0	26.1	9.2	0.806	9	26.9	12.3	0.669	43

HOCCLUS2					
Level	Beta	Row	Column	Density	N
1	0.4	13	10	1.0	60
	0.5	13	10	1.0	60
	0.6	13	10	1.0	60
	0.7	13	10	1.0	60
	0.8	13	10	1.0	60
	0.9	13	10	1.0	60
2	0.4	23.5	19	0.844	30
	0.5	23.5	19	0.844	30
	0.6	23	19	0.855	31
	0.7	23	18.5	0.863	32
	0.8	19	16.5	0.986	40
	0.9	12	11	1.0	53
3	0.4	45	38	0.687	15
	0.5	45	38	0.687	15
	0.6	41.5	33	0.742	18
	0.7	24	19	0.805	25
	0.8	18	17	1.0	35
	0.9	12	10.5	1.0	52
4	0.4	80	63	0.578	8
	0.5	71	58	0.575	9
	0.6	45	40	0.682	13
	0.7	23	19	0.797	22
	0.8	18	17	1.0	34
	0.9	12	10	1.0	51

Example: hsa-let-7c and pluripotency

Let-7 is known to play an essential role in differentiation of embryonic stem cells (ESCs). Sustained expression of let-7 inhibits the reprogramming, and its inhibition promotes the human induced pluripotent stem cell (iPSC) reprogramming²²⁷. PBE algorithm was applied to the let-7c MIR profile, and I found a stem cell specific bicluster comprising 126 target genes and 51 FC conditions (Figure S4.4a and Figure S4.6). This bicluster was quite homogeneous in that 50 of the 51 FC conditions were ESC/iPSC (test) vs. somatic cell (control) conditions suggesting many of the 126 genes are specifically regulated by let-7c or its family microRNAs in stem cells. Indeed, these targets included 49 genes that were reported to have a specific role in ESC (e.g., self-renewal) or upregulated in ESC (Supplementary Table S4.4). Among them, 21 genes (ACVR2B, ARID3B, ²²⁸CCND2, CCNF, CDC25A, DIAPH2, E2F5, HMGA1, IGF2BP1, IGF2BP3, LIN28A, LIN28B, MAPK6, MYCN, PAK1, POU2F1, SERPINB9, SLC5A6, STRBP, USP44 and VAV3) were validated targets of let-7²²⁹⁻²⁴⁴. In particular, MYCN is regulated by let-7 under ESC condition²⁴⁵, and LIN28B is also fine-tuned by let-7 in hESC²²⁸. PLAGL2 which promotes self-renewal in neural stem cell and glioma was also a known target of let-7²⁴⁶. This illustrates the capability of bicluster analysis to identify a specific regulatory module.

Table S4.4. Let-7c bicluster targets regulating pluripotency or up-regulated in ES/iPS cells.

Genes reported to be regulated by let-7 is marked in bold.

Gene symbol	Description	Ref.
ACVR2B	Activin A binds to ActRIIA or ActRIIB and recruits ALK4. ALK4 interacts with SMAD2/3 and activates FGF2 pathways that stimulates self-renewal in human iPS cells by activating target genes including Nanog.	247
ACTA1	Overexpressed in pooled human ES cells compared to huURNA (universal human reference RNA)	248
AMT	Overexpressed in pooled human ES cells compared to huURNA	248
ANKRD46	Ankyrin repeat domain 46. It shows lower CpG methylation and higher gene expression level in pluripotent stem cell compared to somatic cell.	249
ARID3B	ARID3B complex regulates the expression of stemness genes and upregulates the let-7 target genes. Multiple steps in biogenesis of ARID3B-ARID3A complex are regulated by let-7.	250
B3GNT7	The gene expression level of B3GNT7 was 5.84 and 3.16-fold higher in BG02 and BG01 human ES cell line, respectively, compared with the huURNA.	251
C6orf211	Overexpressed in pooled human ES cells compared to huURNA	248
CCND2	CCND2 is a common target of OCT4, SOX2 and NANOG and its overexpression enhances the regenerative potency of hPSC-derived cardiomyocytes.	252-253
CCNF	CCNF (Cyclin F) plays a role in cell cycle event and is essential for embryonic development.	254
CDC25A	NANOG regulates S-phase entry in human ES cells through direct binding of two cell cycle genes CDK6 and CDC25A	255
CDH1	CDH1 regulates open chromatin and pluripotency of embryonic stem cell.	256
CDYL	CDYL is involved in histone modification. It inhibits the neuronal differentiation of iPS cells.	257-258
CTPS2	Overexpressed in pooled human ES cells compared to huURNA	248
DIAPH2	DIAPH2 is involved in actin cytoskeleton pathway and specifically expressed in ES cells.	232, 259
E2F5	E2F4, E2F5 and E2F6 may control E2F target genes during the DNA damage response in human ES cells	260
FZD3	Overexpressed in pooled human ES cells compared to huURNA	248
GALNT13	Overexpressed in pooled human ES cells compared to huURNA	248
GYG2	Overexpressed in pooled human ES cells compared to huURNA	248
HIC2	Overexpressed in pooled human ES cells compared to huURNA	248
HMGA1	HMGA1 is a transcription factor highly expressed in ES cells.	261
HOMER1	Overexpressed in pooled human ES cells compared to huURNA	248
IGF2BP1	IGF2BP1 is highly expressed in ES cells and have important role in human pluripotent stem cell survival.	262
IGF2BP3	IGFBP3 is highly expressed in ES cells compared to differentiated cells.	263
IGSF1	Overexpressed in pooled human ES cells compared to huURNA	248
KIAA1274	Overexpressed in pooled human ES cells compared to huURNA	248

LIN28A	LIN28A regulates mouse iPSC metabolism by let-7-dependent and -independent manner. It is also involved in nucleogenesis during early embryonic development.	264-265
LIN28B	LIN28B have equivalent function with LIN28A	265
MAPK6	Disruption of PI3K/Akt, MAPK/ERK and NFkB signaling pathway results in loss of pluripotency and/or loss of viability. Expression level of MAPK6 was downregulated during the differentiation process.	266
MCM5	MCM5 is involved in DNA replication and up-regulated during the initiation phase of reprogramming.	267
MED28	A mediator subunit, MED28, is required for the acquisition and maintenance of pluripotency during reprogramming	268
MYCN	MYCN maintains embryonic stem cell pluripotency and self-renewal and regulated by let-7.	245, 269
NAP1L1	NAL1L1 regulates the proliferation of murine iPS cells	270
PAK1	PAK1 is involved in actin cytoskeleton pathway and regulates self-renewal activity	232, 271
PLA2G3	Overexpressed in pooled human ES cells compared to huURNA	248
PLAGL2	PLAGL2 promotes self-renewal by regulating Wnt signaling in neural stem cells and glioma	272
POU2F1	Overexpressed in pooled human ES cells compared to huURNA	248
PPP1R16B	PPP1R16B is hypo-methylated and highly expressed in iPS and ES cells	273
RFWD3	Overexpressed in pooled human ES cells compared to huURNA	248
SERPINB9	Overexpressed in pooled human ES cells compared to huURNA	248, 274
SLC16A9	SLC16A9 is a downstream target of OCT4 and upregulated in ES cells.	275
SLC5A6	The gene expression level of SLC5A6 was 3.06 and 3.46-fold higher in BG02 and BG01 hES cell line, respectively, compared with the huRNA (universal human RNA)	251
SMARCD1	SMARCD1 regulates naïve pluripotency by interacting with histone citrullination.	276
SMARCC1	SMARCC1 is involved in chromatin remodeling and highly induced in iPS cells	277
STRBP	Overexpressed in pooled human ES cells compared to huURNA	248
TAF5	TAFs are highly expressed in ES and iPS cells and regulates pluripotency.	278
TARBP2	Overexpressed in pooled human ES cells compared to huURNA	248
TIA1	Overexpressed in pooled human ES cells compared to huURNA	248
THAP9	Overexpressed in pooled human ES cells compared to huURNA	248
USP44	USP44 is highly expressed in ES and IPS cells and it regulates histone H2B ubiquitylation patterns for appropriate ESC differentiation.	279-280
VAV3	Overexpressed in pooled human ES cells compared to huURNA	248

Table S4.5. The accuracy for bicluster targets of eleven test miRNAs

miRNA	1.3-fold			1.5-fold			2.0-fold		
	Sensitivity	Specificity	Gain	Sensitivity	Specificity	Gain	Sensitivity	Specificity	Gain
hsa-miR-1-3p	0.723	0.455	17.81%	0.511	0.559	6.94%	0.489	0.647	13.68%
hsa-miR-21-5p	0.717	0.447	16.41%	0.696	0.57	26.53%	0.674	0.639	31.33%
hsa-miR-125b-5p	0.694	0.5	19.35%	0.371	0.733	10.35%	0.226	0.909	13.47%
hsa-miR-29a-3p	0.727	0.467	19.40%	0.682	0.497	17.88%	0.561	0.656	21.62%
hsa-miR-29b-3p	0.776	0.479	25.48%	0.469	0.726	19.57%	0.531	0.695	22.59%
hsa-miR-29c-3p	0.773	0.438	21.03%	0.75	0.55	30.01%	0.636	0.676	31.28%
hsa-miR-34a-5p	0.73	0.478	20.78%	0.603	0.53	13.34%	0.365	0.707	7.17%
hsa-miR-145-5p	0.7	0.417	11.65%	0.52	0.521	4.11%	0.44	0.651	9.14%
hsa-miR-155-5p	0.556	0.481	3.68%	0.374	0.679	5.29%	0.394	0.637	3.05%
hsa-miR-204-5p	0.737	0.461	19.81%	0.789	0.443	23.26%	0.684	0.619	30.37%
hsa-miR-221-3p	0.639	0.439	7.81%	0.639	0.457	9.61%	0.333	0.649	-1.80%

Table S4.6. The accuracy of 1.3-fold bicluster targets filtered by node degree

Sens.=sensitivity, Spec.=specificity, Gain=Gain in certainty, Overlap P=Overlap p-value, Deplete P=Depletion p-value

	Node degree=1					Node degree=2				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.702	0.582	28.40%	1.53E-03	5.06E-09	0.596	0.678	27.40%	3.46E-03	3.55E-11
hsa-miR-21-5p	0.696	0.582	27.76%	1.47E-03	9.10E-04	0.630	0.725	35.58%	6.21E-05	5.45E-06
hsa-miR-29a-3p	0.712	0.566	27.86%	1.35E-05	4.12E-04	0.621	0.660	28.13%	1.52E-05	1.76E-05
hsa-miR-29b-3p	0.755	0.578	33.35%	1.02E-05	5.05E-06	0.755	0.683	43.85%	7.33E-09	6.58E-10
hsa-miR-29c-3p	0.750	0.549	29.85%	1.06E-04	6.74E-03	0.727	0.675	40.22%	2.40E-07	7.68E-05
hsa-miR-34a-5p	0.714	0.618	33.27%	4.12E-07	5.77E-05	0.619	0.714	33.35%	2.26E-07	2.67E-05
hsa-miR-125b-5p	0.629	0.663	29.18%	2.50E-05	1.93E-06	0.516	0.773	28.94%	6.07E-06	1.35E-05
hsa-miR-145-5p	0.540	0.543	8.31%	1.88E-01	4.72E-02	0.500	0.651	15.14%	2.56E-02	4.68E-02
hsa-miR-155-5p	0.485	0.609	9.39%	2.77E-01	3.38E-05	0.374	0.734	10.81%	2.15E-01	1.56E-06
hsa-miR-204-5p	0.711	0.582	29.21%	4.08E-04	2.16E-02	0.684	0.656	33.99%	4.36E-05	3.76E-03
hsa-miR-221-3p	0.611	0.561	17.19%	2.97E-02	2.69E-01	0.583	0.664	24.77%	2.20E-03	1.54E-01
	Node degree=3					Node degree=4				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.511	0.742	25.24%	4.12E-03	6.93E-09	0.447	0.771	21.75%	7.24E-03	5.25E-06
hsa-miR-21-5p	0.478	0.791	26.92%	3.58E-04	6.72E-03	0.478	0.861	33.89%	1.88E-06	6.59E-04
hsa-miR-29a-3p	0.576	0.727	30.23%	1.68E-06	8.91E-06	0.561	0.782	34.31%	2.24E-08	3.22E-07
hsa-miR-29b-3p	0.714	0.760	47.46%	1.84E-10	5.95E-12	0.694	0.793	48.68%	1.19E-11	8.22E-10
hsa-miR-29c-3p	0.705	0.743	44.72%	4.41E-09	1.62E-05	0.682	0.794	47.53%	1.15E-10	1.23E-05
hsa-miR-34a-5p	0.540	0.795	33.44%	3.46E-08	3.18E-05	0.492	0.842	33.42%	4.71E-09	3.99E-05
hsa-miR-125b-5p	0.468	0.864	33.21%	1.60E-08	6.19E-08	0.371	0.919	28.96%	9.38E-09	1.27E-06
hsa-miR-145-5p	0.420	0.721	14.11%	2.82E-02	6.26E-02	0.320	0.794	11.45%	4.93E-02	6.75E-02
hsa-miR-155-5p	0.333	0.802	13.53%	7.81E-02	1.27E-07	0.303	0.850	15.27%	1.27E-02	1.10E-06
hsa-miR-204-5p	0.632	0.725	35.65%	1.04E-05	4.73E-03	0.553	0.787	34.01%	1.18E-05	2.58E-03
hsa-miR-221-3p	0.556	0.770	32.58%	3.08E-05	5.81E-02	0.389	0.840	22.90%	9.80E-04	1.04E-01
	Node degree=5					Node degree=6				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.362	0.823	18.44%	1.72E-02	4.11E-06	0.362	0.850	21.14%	4.76E-03	6.87E-07
hsa-miR-21-5p	0.457	0.898	35.41%	2.00E-07	1.15E-04	0.391	0.914	30.52%	2.16E-06	3.07E-04
hsa-miR-29a-3p	0.530	0.837	36.72%	3.00E-10	2.97E-08	0.485	0.875	35.95%	8.18E-11	6.09E-09
hsa-miR-29b-3p	0.612	0.845	45.69%	2.43E-11	8.27E-10	0.612	0.880	49.24%	1.23E-13	1.71E-12
hsa-miR-29c-3p	0.659	0.855	51.43%	6.77E-13	3.68E-09	0.614	0.895	50.89%	5.56E-14	8.21E-10
hsa-miR-34a-5p	0.444	0.886	33.00%	6.43E-10	5.55E-06	0.381	0.907	28.75%	8.59E-09	2.01E-05
hsa-miR-125b-5p	0.323	0.952	27.41%	1.41E-09	8.29E-08	0.274	0.969	24.32%	1.20E-09	2.94E-07
hsa-miR-145-5p	0.200	0.855	5.50%	2.15E-01	1.11E-01	0.200	0.901	10.09%	3.71E-02	1.74E-02
hsa-miR-155-5p	0.242	0.882	12.46%	2.39E-02	1.53E-05	0.212	0.920	13.19%	1.71E-02	8.81E-08
hsa-miR-204-5p	0.526	0.827	35.33%	1.88E-06	3.74E-03	0.447	0.873	32.05%	3.68E-06	1.37E-03
hsa-miR-221-3p	0.389	0.860	24.92%	2.59E-04	6.60E-02	0.333	0.896	22.97%	2.66E-04	4.55E-02

Table S4.7. The accuracy of 1.5-fold bicluster targets filtered by node degree

Sens.=sensitivity, Spec.=specificity, Gain=Gain in certainty, Overlap P=Overlap p-value, Deplete P=Depletion p-value

	Node degree=1					Node degree=2				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.426	0.690	11.53%	2.09E-01	5.37E-06	0.383	0.759	14.21%	1.11E-01	7.21E-07
hsa-miR-21-5p	0.609	0.689	29.72%	7.98E-04	1.41E-04	0.522	0.824	34.55%	4.01E-05	1.67E-06
hsa-miR-29a-3p	0.636	0.616	25.27%	9.88E-05	2.02E-04	0.576	0.713	28.87%	6.15E-06	8.08E-06
hsa-miR-29b-3p	0.449	0.812	26.11%	1.44E-04	5.70E-06	0.429	0.864	29.25%	7.79E-06	5.20E-08
hsa-miR-29c-3p	0.727	0.650	37.75%	1.80E-06	2.21E-05	0.659	0.744	40.33%	2.46E-07	2.37E-07
hsa-miR-34a-5p	0.587	0.664	25.18%	8.94E-05	2.42E-03	0.508	0.772	28.03%	5.91E-06	5.88E-05
hsa-miR-125b-5p	0.306	0.822	12.82%	2.67E-02	1.49E-03	0.274	0.886	15.99%	2.09E-03	3.89E-04
hsa-miR-145-5p	0.420	0.631	5.12%	3.17E-01	9.06E-02	0.340	0.741	8.13%	1.49E-01	1.16E-01
hsa-miR-155-5p	0.333	0.784	11.78%	1.04E-01	8.27E-06	0.253	0.840	9.21%	1.28E-01	2.45E-04
hsa-miR-204-5p	0.737	0.583	32.00%	1.11E-04	1.30E-02	0.684	0.675	35.97%	1.16E-05	8.29E-03
hsa-miR-221-3p	0.611	0.579	18.99%	2.38E-02	4.28E-02	0.528	0.703	23.05%	3.41E-03	1.86E-01
	Node degree=3					Node degree=4				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.362	0.811	17.29%	3.29E-02	1.08E-06	0.340	0.840	18.05%	1.59E-02	6.23E-06
hsa-miR-21-5p	0.435	0.873	30.77%	1.99E-05	1.88E-04	0.348	0.918	26.59%	5.97E-05	9.69E-05
hsa-miR-29a-3p	0.530	0.781	31.13%	2.93E-07	4.87E-06	0.515	0.823	33.84%	1.05E-08	9.61E-08
hsa-miR-29b-3p	0.429	0.898	32.65%	1.24E-07	6.38E-09	0.408	0.920	32.83%	1.46E-08	1.57E-08
hsa-miR-29c-3p	0.659	0.803	46.19%	8.74E-10	1.91E-08	0.591	0.841	43.22%	2.62E-09	1.51E-08
hsa-miR-34a-5p	0.444	0.833	27.73%	1.25E-06	1.29E-04	0.397	0.870	26.66%	5.56E-07	2.66E-04
hsa-miR-125b-5p	0.258	0.936	19.41%	3.40E-05	2.96E-06	0.242	0.963	20.51%	4.57E-07	1.62E-06
hsa-miR-145-5p	0.320	0.811	13.10%	2.65E-02	3.31E-02	0.240	0.864	10.42%	4.53E-02	4.27E-02
hsa-miR-155-5p	0.222	0.897	11.95%	4.22E-02	9.91E-07	0.192	0.925	11.67%	1.61E-02	8.24E-06
hsa-miR-204-5p	0.553	0.758	31.05%	6.72E-05	1.86E-02	0.553	0.812	36.48%	1.37E-06	3.80E-03
hsa-miR-221-3p	0.444	0.791	23.50%	1.64E-03	9.20E-02	0.389	0.845	23.35%	7.65E-04	8.56E-02
	Node degree=5					Node degree=6				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.298	0.867	16.49%	1.94E-02	1.22E-05	0.213	0.886	9.91%	1.16E-01	4.03E-04
hsa-miR-21-5p	0.261	0.934	19.53%	8.89E-04	1.52E-03	0.239	0.947	18.59%	5.52E-04	3.40E-03
hsa-miR-29a-3p	0.470	0.867	33.68%	1.87E-09	8.90E-09	0.394	0.893	28.67%	1.46E-08	1.21E-05
hsa-miR-29b-3p	0.388	0.942	33.01%	1.67E-09	5.47E-10	0.347	0.950	29.66%	1.78E-08	8.02E-10
hsa-miR-29c-3p	0.568	0.874	44.18%	2.48E-10	2.69E-09	0.545	0.894	43.91%	4.25E-11	3.11E-08
hsa-miR-34a-5p	0.333	0.896	22.94%	2.14E-06	3.41E-03	0.317	0.921	23.85%	2.08E-07	4.49E-04
hsa-miR-125b-5p	0.194	0.977	17.03%	2.84E-06	1.17E-06	0.129	0.983	11.16%	1.20E-04	1.25E-03
hsa-miR-145-5p	0.200	0.894	9.36%	4.07E-02	1.64E-01	0.160	0.910	7.01%	8.46E-02	2.22E-01
hsa-miR-155-5p	0.182	0.945	12.67%	3.46E-03	3.63E-06	0.131	0.955	8.62%	3.39E-02	5.23E-05
hsa-miR-204-5p	0.474	0.852	32.54%	6.02E-06	1.50E-03	0.368	0.886	25.47%	9.61E-05	5.30E-03
hsa-miR-221-3p	0.333	0.885	21.85%	5.93E-04	8.49E-02	0.306	0.914	22.00%	2.62E-04	2.51E-02

Table S4.8. The accuracy of 2.0-fold bicluster targets filtered by node degree

Sens.=sensitivity, Spec.=specificity, Gain=Gain in certainty, Overlap P=Overlap p-value, Deplete P=Depletion p-value

	Node degree=1					Node degree=2				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.404	0.771	17.50%	2.67E-02	7.89E-05	0.298	0.830	12.83%	9.10E-02	2.55E-05
hsa-miR-21-5p	0.543	0.770	31.40%	8.53E-05	9.21E-04	0.413	0.877	29.01%	8.47E-05	6.60E-05
hsa-miR-29a-3p	0.515	0.745	25.99%	2.47E-05	1.08E-04	0.455	0.810	26.42%	6.33E-06	1.85E-05
hsa-miR-29b-3p	0.510	0.793	30.31%	2.14E-05	3.36E-07	0.469	0.870	33.92%	1.88E-07	1.83E-08
hsa-miR-29c-3p	0.614	0.773	38.71%	3.08E-07	5.47E-06	0.568	0.832	40.02%	2.02E-08	5.91E-06
hsa-miR-34a-5p	0.333	0.804	13.73%	7.25E-03	2.47E-01	0.317	0.851	16.88%	7.00E-04	8.91E-02
hsa-miR-125b-5p	0.210	0.955	16.51%	2.06E-05	6.87E-04	0.177	0.986	16.39%	7.24E-08	2.86E-05
hsa-miR-145-5p	0.340	0.756	9.60%	1.13E-01	2.57E-02	0.300	0.835	13.49%	1.98E-02	1.77E-02
hsa-miR-155-5p	0.333	0.774	10.78%	1.04E-01	1.70E-04	0.273	0.842	11.48%	4.49E-02	1.79E-04
hsa-miR-204-5p	0.605	0.733	33.84%	2.19E-05	1.24E-02	0.474	0.815	28.92%	7.47E-05	2.42E-02
hsa-miR-221-3p	0.306	0.755	6.01%	2.31E-01	7.12E-01	0.278	0.824	10.21%	8.51E-02	4.42E-01
	Node degree=3					Node degree=4				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.298	0.869	16.69%	1.25E-02	1.70E-04	0.255	0.892	14.74%	1.68E-02	5.32E-04
hsa-miR-21-5p	0.391	0.914	30.52%	1.53E-06	5.54E-04	0.283	0.934	21.70%	7.75E-05	8.04E-03
hsa-miR-29a-3p	0.439	0.855	29.44%	1.03E-07	3.09E-06	0.409	0.897	30.64%	2.56E-09	3.23E-07
hsa-miR-29b-3p	0.408	0.904	31.20%	1.40E-07	4.69E-07	0.367	0.942	30.97%	6.33E-09	2.59E-08
hsa-miR-29c-3p	0.523	0.869	39.18%	6.77E-09	5.68E-06	0.455	0.903	35.75%	8.32E-09	3.60E-05
hsa-miR-34a-5p	0.254	0.887	14.08%	1.31E-03	2.53E-01	0.238	0.916	15.39%	1.89E-04	1.44E-01
hsa-miR-125b-5p	0.113	0.994	10.71%	6.55E-06	2.98E-04	0.113	0.996	10.90%	2.13E-06	1.04E-04
hsa-miR-145-5p	0.200	0.879	7.89%	1.04E-01	3.55E-02	0.200	0.916	11.56%	1.43E-02	1.24E-02
hsa-miR-155-5p	0.242	0.910	15.22%	2.08E-03	2.83E-06	0.182	0.935	11.67%	4.21E-03	1.93E-04
hsa-miR-204-5p	0.395	0.888	28.27%	1.47E-05	8.50E-03	0.316	0.918	23.34%	7.18E-05	1.24E-02
hsa-miR-221-3p	0.250	0.876	12.61%	2.49E-02	5.03E-01	0.250	0.919	16.89%	2.39E-03	1.13E-01
	Node degree=5					Node degree=6				
miRNA	Sens.	Spec.	Gain	Overlap P	Deplete P	Sens.	Spec.	Gain	Overlap P	Deplete P
hsa-miR-1-3p	0.213	0.929	14.15%	1.23E-02	9.00E-05	0.170	0.950	12.01%	1.54E-02	2.87E-04
hsa-miR-21-5p	0.196	0.955	15.06%	1.44E-03	2.13E-02	0.109	0.975	8.41%	1.19E-02	1.60E-01
hsa-miR-29a-3p	0.348	0.921	26.99%	8.85E-09	3.54E-06	0.318	0.944	26.23%	1.64E-09	5.97E-07
hsa-miR-29b-3p	0.347	0.957	30.40%	2.20E-09	1.06E-09	0.327	0.963	28.95%	1.47E-09	3.34E-08
hsa-miR-29c-3p	0.432	0.928	35.94%	8.15E-10	8.18E-06	0.432	0.948	37.94%	8.08E-12	8.13E-07
hsa-miR-34a-5p	0.206	0.936	14.19%	2.08E-04	7.75E-02	0.190	0.950	14.05%	1.02E-04	2.99E-02
hsa-miR-125b-5p	0.048	0.996	4.45%	8.10E-03	3.87E-02	0.048	0.998	4.64%	3.49E-03	1.77E-02
hsa-miR-145-5p	0.180	0.938	11.76%	5.73E-03	2.31E-02	0.140	0.952	9.23%	1.47E-02	3.42E-02
hsa-miR-155-5p	0.131	0.957	8.87%	1.70E-02	1.86E-04	0.111	0.972	8.35%	1.53E-02	5.06E-05
hsa-miR-204-5p	0.237	0.946	18.25%	3.81E-04	9.29E-03	0.158	0.969	12.66%	2.78E-03	9.28E-03
hsa-miR-221-3p	0.222	0.941	16.37%	1.29E-03	9.01E-02	0.194	0.968	16.29%	3.04E-04	1.49E-02

Table S4.9. microRNA expression patterns in cancers reported from the literature

microRNA	Cancer	Direction	Reference
hsa-miR-1-3p	Breast cancer	Down-regulation	281
hsa-miR-21-5p	Breast cancer	Up-regulation	282
hsa-miR-29a-3p	Acute Myeloid Leukemia	Down-regulation	283-284
	Breast cancer	Down-regulation	285
	Diffuse Large B-cell Lymphoma	Down-regulation	286
	Glioblastoma/glioma	Down-regulation	287
hsa-miR-29b-3p	Acute Myeloid Leukemia	Down-regulation	288
	Breast cancer	Down-regulation	289
	Diffuse Large B-cell Lymphoma	Down-regulation	286
	Glioblastoma/glioma	Down-regulation	290
hsa-miR-29c-3p	Breast cancer	Down-regulation	291
	Diffuse Large B-cell Lymphoma	Down-regulation	286
	Glioblastoma/glioma	Down-regulation	290
hsa-miR-34a-5p	Breast cancer	Down-regulation	292
	Diffuse Large B-cell Lymphoma	Down-regulation	293
hsa-miR-125a-5p	Acute Myeloid Leukemia	Up-regulation	294
hsa-miR-145-5p	Acute Myeloid Leukemia	Down-regulation	295
	Breast cancer	Down-regulation	296
	Diffuse Large B-cell Lymphoma	Down-regulation	297
hsa-miR-155-5p	Breast cancer	Up-regulation	298
hsa-miR-221-3p	Breast cancer	Up-regulation	299

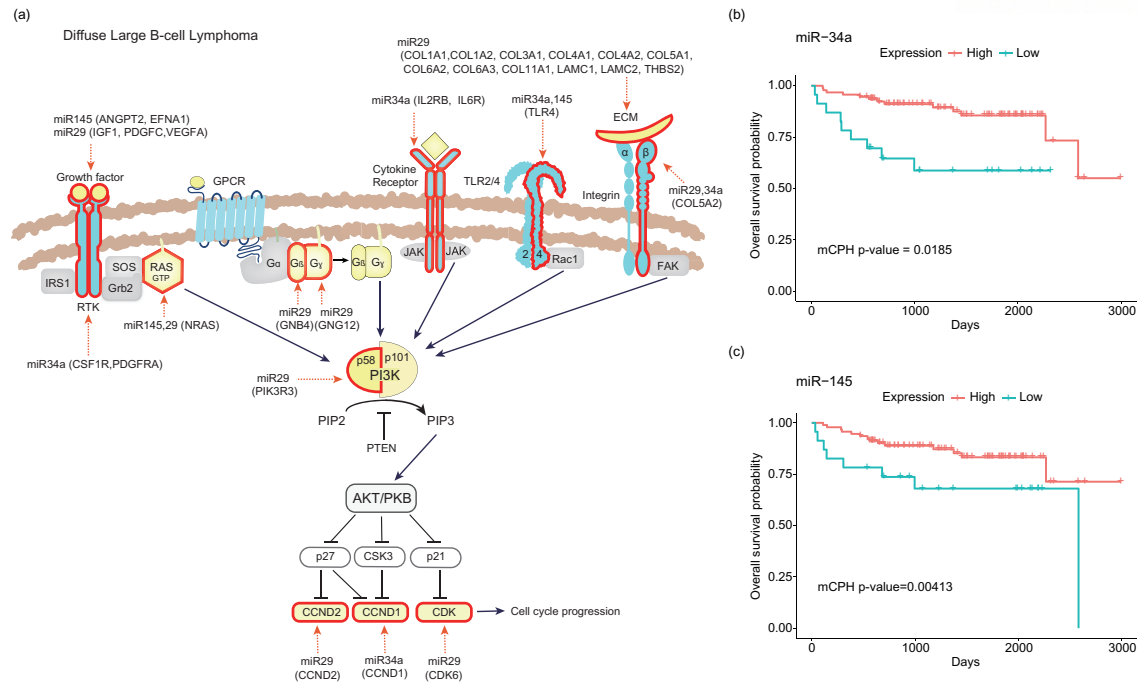


Figure S4.5. microRNA targets in PI3K/Akt pathway (DLBCL).

(a) MicroRNA targets predicted from DLBCL biclusters in PI3K/Akt pathway are highlighted by red borders. For each target molecule, corresponding microRNAs and target gene symbols are represented.

(b, c) Overall survival analysis for the 116 DLBCL patients (GSE40239) of high (red) and low (blue) expression levels. The patients were divided into two groups based on their best splits (both at bottom 20% values).

Table S4.10. Functional enrichment test for miR-29, miR-34a, miR-145 targets in DLBCL

Term	Count	P-value	FDR
hsa04151:PI3K-Akt signaling pathway	31	4.19E-10	9.09E-08
hsa04510:Focal adhesion	22	1.46E-08	1.59E-06
hsa04512:ECM-receptor interaction	13	8.66E-07	6.26E-05
hsa04974:Protein digestion and absorption	13	9.82E-07	5.33E-05
hsa05146:Amoebiasis	12	4.03E-05	1.75E-03
hsa04014:Ras signaling pathway	17	9.00E-05	3.25E-03
hsa05218:Melanoma	8	1.44E-03	4.35E-02
hsa05222:Small cell lung cancer	8	4.03E-03	1.04E-01
hsa04015:Rap1 signaling pathway	13	4.31E-03	9.89E-02
hsa05214:Glioma	7	4.37E-03	9.06E-02
hsa05162:Measles	10	4.40E-03	8.33E-02
hsa04066:HIF-1 signaling pathway	8	8.74E-03	1.47E-01
hsa04060:Cytokine-cytokine receptor interaction	13	8.76E-03	1.37E-01
hsa05200:Pathways in cancer	18	1.23E-02	1.74E-01
hsa05215:Prostate cancer	7	1.83E-02	2.35E-01
hsa05212:Pancreatic cancer	6	1.91E-02	2.30E-01
hsa04115:p53 signaling pathway	6	2.15E-02	2.42E-01
hsa04360:Axon guidance	8	3.19E-02	3.24E-01
hsa04611:Platelet activation	8	3.56E-02	3.39E-01
hsa04668:TNF signaling pathway	7	4.08E-02	3.64E-01
hsa05223:Non-small cell lung cancer	5	4.41E-02	3.72E-01
hsa04144:Endocytosis	12	4.43E-02	3.60E-01
hsa04150:mTOR signaling pathway	5	4.91E-02	3.78E-01
hsa05205:Proteoglycans in cancer	10	4.96E-02	3.69E-01
hsa04550:Signaling pathways regulating pluripotency of stem cells	8	4.98E-02	3.58E-01
hsa04540:Gap junction	6	5.88E-02	3.97E-01
hsa05219:Bladder cancer	4	7.37E-02	4.59E-01
hsa04110:Cell cycle	7	7.58E-02	4.57E-01
hsa05166:HTLV-I infection	11	8.54E-02	4.87E-01
hsa05220:Chronic myeloid leukemia	5	9.27E-02	5.05E-01
hsa05231:Choline metabolism in cancer	6	9.34E-02	4.97E-01

Table S4.11. Functional enrichment test for miR-1, miR-29, miR-34a, miR-145 targets in breast cancer

Term	Count	P-value	FDR
hsa04510:Focal adhesion	23	7.33E-12	1.38E-09
hsa04512:ECM-receptor interaction	16	2.15E-11	2.02E-09
hsa04151:PI3K-Akt signaling pathway	27	2.11E-10	1.32E-08
hsa05146:Amoebiasis	12	2.42E-06	1.14E-04
hsa04974:Protein digestion and absorption	11	3.09E-06	1.16E-04
hsa05200:Pathways in cancer	21	1.71E-05	5.35E-04
hsa05222:Small cell lung cancer	10	1.77E-05	4.74E-04
hsa05205:Proteoglycans in cancer	12	8.22E-04	1.91E-02
hsa04611:Platelet activation	9	2.13E-03	4.35E-02
hsa05219:Bladder cancer	5	5.82E-03	1.04E-01
hsa04360:Axon guidance	8	7.32E-03	1.18E-01
hsa05218:Melanoma	6	8.41E-03	1.24E-01
hsa05215:Prostate cancer	6	2.00E-02	2.53E-01
hsa05212:Pancreatic cancer	5	2.81E-02	3.18E-01
hsa05166:HTLV-I infection	10	3.84E-02	3.88E-01
hsa05220:Chronic myeloid leukemia	5	3.89E-02	3.73E-01
hsa05161:Hepatitis B	7	4.43E-02	3.94E-01
hsa04014:Ras signaling pathway	9	4.79E-02	4.01E-01
hsa00532:Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	3	4.85E-02	3.88E-01
hsa05144:Malaria	4	5.61E-02	4.19E-01
hsa05145:Toxoplasmosis	6	5.89E-02	4.19E-01
hsa04152:AMPK signaling pathway	6	6.60E-02	4.42E-01
hsa05223:Non-small cell lung cancer	4	7.73E-02	4.82E-01
hsa05202:Transcriptional misregulation in cancer	7	7.87E-02	4.74E-01
hsa04150:mTOR signaling pathway	4	8.39E-02	4.83E-01
hsa04010:MAPK signaling pathway	9	8.41E-02	4.70E-01
hsa04915:Estrogen signaling pathway	5	9.95E-02	5.18E-01

Table S4.12. Multivariate Cox regression analysis of microRNAs in the DLBCL dataset

Variable	Hazard ratio	95% CI. *	p-value
<i>miR-29a</i>			
miR-29a	0.903	0.662-1.231	5.18.E-01
IPI *	1.720	1.231-2.404	1.48.E-03
Gender	2.498	1.043-5.982	3.98.E-02
<i>miR-29b</i>			
miR-29b	0.912	0.664-1.252	5.68.E-01
IPI	1.751	1.27-2.412	6.20.E-04
Gender	2.558	1.063-6.154	3.61.E-02
<i>miR-29c</i>			
miR-29c	0.833	0.581-1.193	3.19.E-01
IPI	1.721	1.242-2.386	1.12.E-03
Gender	2.609	1.084-6.277	3.23.E-02
<i>miR-34a</i>			
miR-34a	0.691	0.508-0.94	1.85.E-02
IPI	1.687	1.225-2.322	1.35.E-03
Gender	2.983	1.171-7.6	2.20.E-02
<i>miR-145</i>			
miR-145	0.593	0.415-0.848	4.13.E-03
IPI	1.787	1.312-2.434	2.28.E-04
Gender	3.075	1.266-7.466	1.31.E-02

*CI=Confidence Interval, IPI=International prognostic index

Table S4.13. Multivariate Cox regression analysis of microRNAs in the breast cancer dataset

	Hazard ratio	95% CI. *	p-value
<i>miR-1</i>			
miR-1	1.034	0.848-1.261	7.43.E-01
Age	1.036	1.011-1.061	4.22.E-03
Tumor size	1.206	1.018-1.43	3.08.E-02
Lymph nodes involved	1.194	1.125-1.268	6.13.E-09
ER *	0.666	0.396-1.12	1.25.E-01
Grade	1.603	1.099-2.339	1.43.E-02
<i>miR-29a</i>			
miR-29a	0.745	0.609-0.911	4.22.E-03
Age	1.039	1.014-1.065	2.36.E-03
Tumor size	1.213	1.03-1.428	2.05.E-02
Lymph nodes involved	1.213	1.141-1.289	5.70.E-10
ER	0.605	0.356-1.027	6.28.E-02
Grade	1.477	1.012-2.156	4.29.E-02
<i>miR-29b</i>			
miR-29b	0.717	0.565-0.911	6.42.E-03
Age	1.041	1.016-1.067	1.02.E-03
Tumor size	1.245	1.058-1.465	8.40.E-03
Lymph nodes involved	1.209	1.136-1.287	2.08.E-09
ER	0.713	0.424-1.2	2.03.E-01
Grade	1.712	1.176-2.49	4.96.E-03
<i>miR-29c</i>			
miR-29c	0.715	0.57-0.897	3.80.E-03
Age	1.037	1.014-1.061	1.75.E-03
Tumor size	1.256	1.064-1.483	7.15.E-03
Lymph nodes involved	1.195	1.127-1.267	2.06.E-09
ER	0.796	0.467-1.356	4.01.E-01
Grade	1.469	1.012-2.13	4.29.E-02
<i>miR-34a</i>			
miR-34a	1.023	0.795-1.316	8.62.E-01
Age	1.036	1.011-1.062	4.10.E-03
Tumor size	1.216	1.034-1.429	1.80.E-02
Lymph nodes involved	1.193	1.124-1.266	5.43.E-09
ER	0.669	0.398-1.124	1.29.E-01
Grade	1.573	1.083-2.286	1.75.E-02
<i>miR-145</i>			
miR-145	1.168	0.921-1.482	2.00.E-01
Age	1.038	1.013-1.063	2.92.E-03
Tumor size	1.224	1.042-1.437	1.37.E-02
Lymph nodes involved	1.195	1.126-1.267	2.90.E-09
ER	0.702	0.418-1.181	1.83.E-01
Grade	1.631	1.126-2.362	9.69.E-03

*CI=Confidence Interval, ER=Estrogen receptor

MicroRNA regulation of PI3K/Akt pathways in the literature

The microRNAs detected in cancer biclusters were able to suppress PI3K/Akt pathway and metastasis in multiple cancer types. For example, up-regulated miR-29a inhibited the lung cancer proliferation by targeting NRAS which is a key downstream effector of PI3K/Akt pathway³⁰⁰. Up-regulated MiR-29b suppressed the breast cancer metastasis by targeting VEGFA, PDGFC and ITGB1²¹⁷, and it also reduced angiogenesis of endometrial cancer by targeting VEGFA³⁰¹. MiR-34a inhibited gastric cancer growth, invasion, and metastasis by targeting two signal transducers of the pathway, PDGFR and MET³⁰². Mir-1 also acted as a tumor suppressor in gastric cancer by targeting VEGFA and MET³⁰³⁻³⁰⁴. Lastly, miR-145 inhibited PI3K/Akt pathway by targeting NRAS in melanoma³⁰⁵. The same targets and microRNAs were detected in our bicluster results for breast cancer and DLBCL, suggesting these microRNAs are also able to suppress PI3K/Akt pathway and metastasis in these cancer types. Indeed, it was shown in vivo that mir-29b considerably inhibits breast cancer metastasis by suppressing tumor microenvironment related targets²¹⁷. Our biclustering result suggests collagen and other genes in PI3K/Akt pathway are also targets of mir-29 in breast cancer and DLBCL.

miRNA mimic transfection assays

miR-29c-3p mimic and miRNA scramble control were purchased from Genolution. 100nM of miR-29c-3p mimic and miRNA scramble were transfected into MDA-MB231 using G-fectin Reagent (Genolution). Experiments were performed 48 hours after transfection.

Real-Time Quantitative PCR

One microgram of total RNA from MDA-MB231 cell was reverse transcribed with oligo dT and M-MLV RT reverse transcriptase (Invitrogen). Real-time quantitative PCR was performed using a GENETBIO SYBR Green Prime Q-master Mix and the QuantStudio 5 PCR system (ThermoFisher). All runs were accompanied by the internal control HPRT or B2M gene. The samples were run in duplicate and normalized to HPRT or GAPDH using a DD cycle threshold-based algorithm, to provide arbitrary units representing relative expression.

immunoblot assays

For immunoblot assay, cells were lysed in RIPA buffer. Protein concentrations were determined with the BCA Protein Assay (ThermoFisher). Quantified lysates were loaded on SDS-PAGE, transferred onto NC membrane and probed with rabbit anti-human FAK (1:1000, cell signaling), phosphor FAK (1:1000, cell signaling), Akt (1:1000, cell signaling), phosphor Akt (1:1000, cell signaling) and mouse anti-human GAPDH (1:1000, cell signaling) followed by incubation with secondary fluorescent antibodies (1:5000, Li-COR).

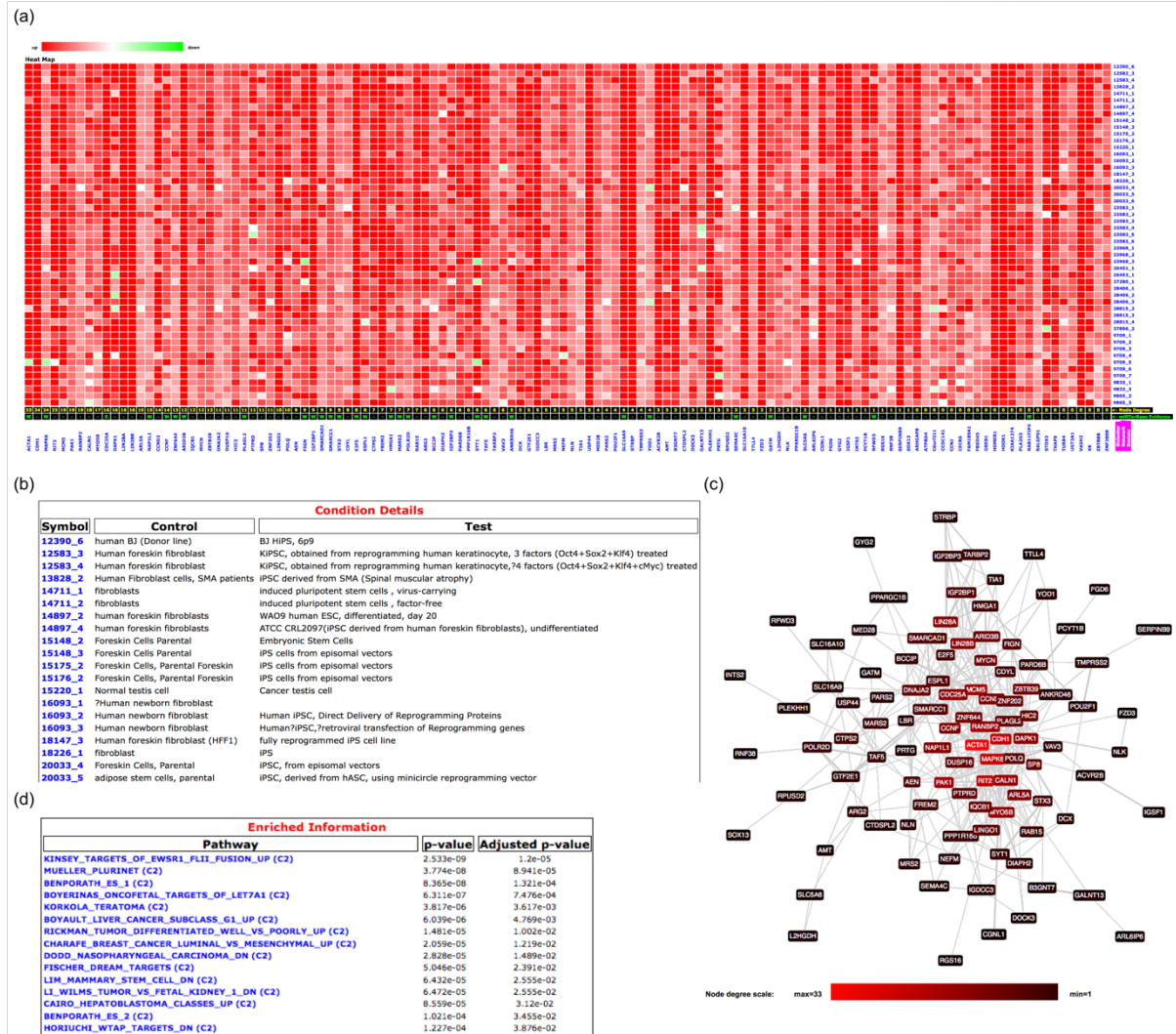


Figure S4.6. BiMIR database.

(a) Heatmap of hsa-let-7c-5p bicluster up-regulated under embryonic stem cell/iPS cell/somatic cell conditions. Row and column represent the symbols of experimental conditions and target gene symbols, respectively. For each target gene, the user can check the node-degree for target PPI network and whether it is experimentally validated. (b) Detailed condition information (test and control group info.) is represented. Wordcloud for conditions is also provided. (c) PPI network for bicluster targets are visualized. The nodes with bright red color are connected with many other targets.

Chapter V: Discussion and conclusion

In this dissertation, the algorithms to improve the pathway analysis of RNA-seq data with small replicates and GWAS summary data were addressed, as well as an approach to predict cell condition-specific miRNA targets by biclustering the big transcriptomic data.

In Chapter II, the effect of absolute statistic in reducing the false positive results from gene-permuting GSEA was confirmed through simulation and real data analysis. It was observed that the absolute statistic reduced the variance inflation factor when tested with TCGA cancer data. Based on this observation, I developed an R package named ‘AbsFilterGSEA’. It provides a useful function that filters significant gene-sets detected by original gene-permuting GSEA with those detected by absolute GSEA, so that users obtain reliable gene-sets with known directionality. However, the reason why absolute statistic relieves the variance inflation factor is still in question and it should be mathematically proved through further study.

In Chapter III, it was observed that the z-score method applied with modified gene scores (adjusted by SNP size) greatly improved sensitivity compared to existing competitive and some of self-contained gene-set analysis methods, while exhibiting decent false positive control. Also, it was good at prioritizing phenotype-related pathways, and showed outstanding performance when the sample size is relatively small (KARE data, < 9000 samples). In addition, it provides gene network visualization within a gene-set or across significant gene-sets. From the global network (across significant gene-sets), users can identify hub gene or core sub-network that may affect to multiple pathways, and thus plays a central role in corresponding disease. Currently, it provides only two PPI database (STRING and HIPPIE). In the update version, more PPI sources need to be included. Also, using the extended PPI network, it might be possible to infer the biological role of unannotated genes showing meaningful association signals.

In Chapter IV, a novel approach to predict condition-specific miRNA target network by biclustering the transcriptomic big data was addressed. Compared to pure sequence-based method, biclustering improved the target prediction and the accuracy was further improved by filtering the bicluster targets using network information. In addition, the bicluster targets were favorable when compared with targets predicted from TCGA mRNA-miRNA paired expression data. The cancer bicluster analysis revealed that few miRNAs’ targets were enriched in ‘PI3K/Akt signaling pathway’, and it was experimentally validated that miR-29 suppresses nine genes involved in the pathway. There are two future works for this project. First, I can also apply this approach to other types of regulators that binds to DNA in a sequence-specific manner such as transcription factor. Second, it is possible to construct and infer miRNA co-regulation network using significantly overlapping biclusters.

References

1. International Human Genome Sequencing, C., Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431* (7011), 931-45.
2. Frazer, K. A.; Ballinger, D. G.; Cox, D. R.; Hinds, D. A.; Stuve, L. L.; Gibbs, R. A.; Belmont, J. W.; Boudreau, A.; Hardenbol, P.; Leal, S. M.; Pasternak, S.; Wheeler, D. A.; Willis, T. D.; Yu, F. L.; Yang, H. M.; Zeng, C. Q.; Gao, Y.; Hu, H. R.; Hu, W. T.; Li, C. H.; Lin, W.; Liu, S. Q.; Pan, H.; Tang, X. L.; Wang, J.; Wang, W.; Yu, J.; Zhang, B.; Zhang, Q. R.; Zhao, H. B.; Zhao, H.; Zhou, J.; Gabriel, S. B.; Barry, R.; Blumenstiel, B.; Camargo, A.; Defelice, M.; Faggart, M.; Goyette, M.; Gupta, S.; Moore, J.; Nguyen, H.; Onofrio, R. C.; Parkin, M.; Roy, J.; Stahl, E.; Winchester, E.; Ziaugra, L.; Altshuler, D.; Shen, Y.; Yao, Z. J.; Huang, W.; Chu, X.; He, Y. G.; Jin, L.; Liu, Y. F.; Shen, Y. Y.; Sun, W. W.; Wang, H. F.; Wang, Y.; Wang, Y.; Xiong, X. Y.; Xu, L.; Waye, M. M. Y.; Tsui, S. K. W.; Wong, J. T. F.; Galver, L. M.; Fan, J. B.; Gunderson, K.; Murray, S. S.; Oliphant, A. R.; Chee, M. S.; Montpetit, A.; Chagnon, F.; Ferretti, V.; Leboeuf, M.; Olivier, J. F.; Phillips, M. S.; Roumy, S.; Sallee, C.; Verner, A.; Hudson, T. J.; Kwok, P. Y.; Cai, D. M.; Koboldt, D. C.; Miller, R. D.; Pawlikowska, L.; Taillon-Miller, P.; Xiao, M.; Tsui, L. C.; Mak, W.; Song, Y. Q.; Tam, P. K. H.; Nakamura, Y.; Kawaguchi, T.; Kitamoto, T.; Morizono, T.; Nagashima, A.; Ohnishi, Y.; Sekine, A.; Tanaka, T.; Tsunoda, T.; Deloukas, P.; Bird, C. P.; Delgado, M.; Dermitzakis, E. T.; Gwilliam, R.; Hunt, S.; Morrison, J.; Powell, D.; Stranger, B. E.; Whittaker, P.; Bentley, D. R.; Daly, M. J.; de Bakker, P. I. W.; Barrett, J.; Chretien, Y. R.; Maller, J.; McCarroll, S.; Patterson, N.; Pe'er, I.; Price, A.; Purcell, S.; Richter, D. J.; Sabeti, P.; Saxena, R.; Schaffner, S. F.; Sham, P. C.; Varilly, P.; Altshuler, D.; Stein, L. D.; Krishnan, L.; Smith, A. V.; Tello-Ruiz, M. K.; Thorisson, G. A.; Chakravarti, A.; Chen, P. E.; Cutler, D. J.; Kashuk, C. S.; Lin, S.; Abecasis, G. R.; Guan, W. H.; Li, Y.; Munro, H. M.; Qin, Z. H. S.; Thomas, D. J.; McVean, G.; Auton, A.; Bottolo, L.; Cardin, N.; Eyheramendy, S.; Freeman, C.; Marchini, J.; Myers, S.; Spencer, C.; Stephens, M.; Donnelly, P.; Cardon, L. R.; Clarke, G.; Evans, D. M.; Morris, A. P.; Weir, B. S.; Tsunoda, T.; Johnson, T. A.; Mullikin, J. C.; Sherry, S. T.; Feolo, M.; Skol, A.; Consortium, I. H., A second generation human haplotype map of over 3.1 million SNPs. *Nature* **2007**, *449* (7164), 851-U3.
3. Altshuler, D. M.; Durbin, R. M.; Abecasis, G. R.; Bentley, D. R.; Chakravarti, A.; Clark, A. G.; Donnelly, P.; Eichler, E. E.; Flicek, P.; Gabriel, S. B.; Gibbs, R. A.; Green, E. D.; Hurler, M. E.; Knoppers, B. M.; Korb, J. O.; Lander, E. S.; Lee, C.; Lehrach, H.; Mardis, E. R.; Marth, G. T.; McVean, G. A.; Nickerson, D. A.; Schmidt, J. P.; Sherry, S. T.; Wang, J.; Wilson, R. K.; Gibbs, R. A.; Boerwinkle, E.; Doddapaneni, H.; Han, Y.; Korchina, V.; Kovar, C.; Lee, S.; Muzny, D.; Reid, J. G.; Zhu, Y. M.; Wang, J.; Chang, Y. Q.; Feng, Q.; Fang, X. D.; Guo, X. S.; Jian, M.; Jiang, H.; Jin, X.; Lan, T. M.; Li, G. Q.; Li, J. X.; Li, Y. R.; Liu, S. M.; Liu, X.; Lu, Y.; Ma, X. D.; Tang, M. F.; Wang, B.; Wang, G. B.; Wu, H. L.; Wu, R. H.; Xu, X.; Yin, Y.; Zhang, D. D.; Zhang, W. W.; Zhao, J.; Zhao, M. R.; Zheng, X. L.; Lander, E. S.; Altshuler, D. M.; Gabriel, S. B.; Gupta, N.; Gharani, N.; Toji, L. H.; Gerry, N. P.; Resch, A. M.; Flicek, P.; Barker, J.; Clarke, L.; Gil, L.; Hunt, S. E.; Kelman, G.; Kulesha, E.; Leinonen, R.; McLaren, W. M.; Radhakrishnan, R.; Roa, A.; Smirnov, D.; Smith, R. E.; Streeter, I.; Thormann, A.; Toneva, I.; Vaughan, B.; Zheng-Bradley, X.; Bentley, D. R.; Grocock, R.; Humphray, S.; James, T.; Kingsbury, Z.; Lehrach, H.; Sudbrak, R.; Albrecht, M. W.; Amstislavskiy, V. S.; Borodina, T. A.; Lienhard, M.; Mertes, F.; Sultan, M.; Timmermann, B.; Yaspo, M. L.; Mardis, E. R.; Wilson, R. K.; Fulton, R.; Fulton, R.; Sherry, S. T.; Ananiev, V.; Belaia, Z.; Beloslyudtsev, D.; Bouk, N.; Chen, C.; Church, D.; Cohen, R.; Cook, C.; Garner, J.; Heffernon, T.; Kimelman, M.; Liu, C. L.; Lopez, J.; Meric, P.; O'Sullivan, C.; Ostapchuk, Y.; Phan, L.; Ponomarev, S.; Schneider, V.; Shekhtman, E.; Sirotkin, K.; Slotta, D.; Zhang, H.; McVean, G. A.; Durbin, R. M.; Balasubramaniam, S.; Burton, J.; Danecek, P.; Keane, T. M.; Kolb-Kocinski, A.; McCarthy, S.; Stalker, J.; Quail, M.; Schmidt, J. P.; Davies, C. J.; Gollub, J.; Webster, T.; Wong, B.; Zhan, Y. P.; Auton, A.; Campbell, C. L.; Kong, Y.; Marcketta, A.; Gibbs, R. A.; Yu, F. L.; Antunes, L.; Bainbridge, M.; Muzny, D.; Sabo, A.; Huang, Z. Y.; Wang, J.; Coin, L. J. M.; Fang, L.; Guo, X. S.; Jin, X.; Li, G. Q.; Li, Q. B.; Li, Y. R.; Li, Z. Y.; Lin, H. X.; Liu, B. H.; Luo, R. B.; Shao, H. J.; Xie, Y. L.; Ye, C.; Yu, C.; Zhang, F.; Zheng, H. C.; Zhu, H. M.; Alkan, C.; Dal, E.; Kahveci, F.; Marth, G. T.; Garrison, E. P.; Kural, D.; Lee, W. P.; Leong, W. F.; Stromberg, M.; Ward, A. N.; Wu, J. T.; Zhang, M. Y.; Daly, M. J.; DePristo, M. A.; Handsaker, R. E.; Altshuler, D. M.; Banks, E.; Bhatia, G.; del Angel, G.; Gabriel, S. B.; Genovese, G.; Gupta, N.; Li, H.; Kashin, S.; Lander, E. S.; McCarroll, S. A.; Nemesh, J. C.; Poplin, R. E.; Yoon, S. C.; Lihm, J.; Makarov, V.; Clark, A. G.; Gottipati, S.; Keinan, A.; Rodriguez-Flores, J. L.; Korb, J. O.; Rausch, T.; Fritz, M. H.; Stuetz, A. M.; Flicek, P.; Beal, K.; Clarke, L.; Datta, A.; Herrero, J.; McLaren, W. M.; Ritchie, G. R. S.; Smith, R. E.; Zerbino, D.; Zheng-Bradley, X.; Sabeti, P. C.; Shlyakhter, I.; Schaffner, S. F.; Vitti, J.; Cooper, D. N.; Ball, E. V.; Stenson, P. D.; Bentley, D. R.; Barnes, B.; Bauer, M.; Cheetham, R. K.; Cox, A.; Eberle, M.; Humphray, S.; Kahn, S.; Murray, L.; Peden, J.; Shaw, R.; Kenny, E. E.; Batzer, M. A.; Konkel, M. K.; Walker, J. A.; MacArthur, D. G.; Lek, M.; Sudbrak, R.; Amstislavskiy,

V. S.; Herwig, R.; Mardis, E. R.; Ding, L.; Koboldt, D. C.; Larson, D.; Ye, K.; Gravel, S.; Swaroop, A.; Chew, E.; Lappalainen, T.; Erlich, Y.; Gymrek, M.; Willems, T. F.; Simpson, J. T.; Shriver, M. D.; Rosenfeld, J. A.; Bustamante, C. D.; Montgomery, S. B.; De La Vega, F. M.; Byrnes, J. K.; Carroll, A. W.; DeGorter, M. K.; Lacroute, P.; Maples, B. K.; Martin, A. R.; Moreno-Estrada, A.; Shringarpure, S. S.; Zakharia, F.; Halperin, E.; Baran, Y.; Lee, C.; Cerveira, E.; Hwang, J.; Malhotra, A.; Plewczynski, D.; Radew, K.; Romanovitch, M.; Zhang, C. S.; Hyland, F. C. L.; Craig, D. W.; Christoforides, A.; Homer, N.; Izatt, T.; Kurdoglu, A. A.; Sinari, S. A.; Squire, K.; Sherry, S. T.; Xiao, C. L.; Sebat, J.; Antaki, D.; Gujral, M.; Noor, A.; Ye, K.; Burchard, E. G.; Hernandez, R. D.; Gignoux, C. R.; Haussler, D.; Katzman, S. J.; Kent, W. J.; Howie, B.; Ruiz-Linares, A.; Dermitzakis, E. T.; Devine, S. E.; Goncalo, R. A.; Kang, H. M.; Kidd, J. M.; Blackwell, T.; Caron, S.; Chen, W.; Emery, S.; Fritsche, L.; Fuchsberger, C.; Jun, G.; Li, B. S.; Lyons, R.; Scheller, C.; Sidore, C.; Song, S. Y.; Sliwerska, E.; Taliun, D.; Tan, A.; Welch, R.; Wing, M. K.; Zhan, X. W.; Awadalla, P.; Hodgkinson, A.; Li, Y.; Shi, X. H.; Quitadamo, A.; Lunter, G.; McVean, G. A.; Marchini, J. L.; Myers, S.; Churchhouse, C.; Delaneau, O.; Gupta-Hinch, A.; Kretzschmar, W.; Iqbal, Z.; Mathieson, I.; Menelaou, A.; Rimmer, A.; Xifara, D. K.; Oleksyk, T. K.; Fu, Y. X.; Liu, X. M.; Xiong, M. M.; Jorde, L.; Witherspoon, D.; Xing, J. C.; Eichler, E. E.; Browning, B. L.; Browning, S. R.; Hormozdiari, F.; Sudmant, P. H.; Khurana, E.; Durbin, R. M.; Hurles, M. E.; Tyler-Smith, C.; Albers, C. A.; Ayub, Q.; Balasubramaniam, S.; Chen, Y.; Colonna, V.; Danecek, P.; Jostins, L.; Keane, T. M.; McCarthy, S.; Walter, K.; Xue, Y. L.; Gerstein, M. B.; Abyzov, A.; Balasubramanian, S.; Chen, J. M.; Clarke, L.; Fu, Y.; Harmanci, A. O.; Jin, M.; Lee, D.; Liu, J.; Mu, X. J.; Zhang, J.; Zhang, Y.; Li, Y. R.; Luo, R. B.; Zhu, H. M.; Alkan, C.; Dal, E.; Kahveci, F.; Marth, G. T.; Garrison, E. P.; Kural, D.; Lee, W. P.; Ward, A. N.; Wu, J. T.; Zhang, M. Y.; McCarroll, S. A.; Handsaker, R. E.; Altshuler, D. M.; Banks, E.; Del Angel, G.; Genovese, G.; Hartl, C.; Li, H.; Kashin, S.; Nemesh, J. C.; Shakir, K.; Yoon, S. C.; Lihm, J.; Makarov, V.; Degenhardt, J.; Korbel, J. O.; Fritz, M. H.; Meiers, S.; Raeder, B.; Rausch, T.; Stuetz, A. M.; Flicek, P.; Casale, F. P.; Clarke, L.; Smith, R. E.; Stegle, O.; Zheng-Bradley, X.; Bentley, D. R.; Barnes, B.; Cheetham, R. K.; Eberle, M.; Humphray, S.; Kahn, S.; Murray, L.; Shaw, R.; Lameijer, E. W.; Batzer, M. A.; Konkel, M. K.; Walker, J. A.; Ding, L.; Hall, I.; Ye, K.; Lacroute, P.; Lee, C.; Cerveira, E.; Malhotra, A.; Hwang, J.; Plewczynski, D.; Radew, K.; Romanovitch, M.; Zhang, C. S.; Craig, D. W.; Homer, N.; Church, D.; Xiao, C. L.; Sebat, J.; Antaki, D.; Bafna, V.; Michaelson, J.; Ye, K.; Devine, S. E.; Gardner, E. J.; Abecasis, G. R.; Kidd, J. M.; Mills, R. E.; Dayama, G.; Emery, S.; Jun, G.; Shi, X. H.; Quitadamo, A.; Lunter, G.; McVean, G. A.; Chen, K.; Fan, X.; Chong, Z. C.; Chen, T. H.; Witherspoon, D.; Xing, J. C.; Eichler, E. E.; Chaisson, M. J.; Hormozdiari, F.; Huddleston, J.; Malig, M.; Nelson, B. J.; Sudmant, P. H.; Parrish, N. F.; Khurana, E.; Hurles, M. E.; Ben B. l. a. c. k. b. u. r. n. e. ; Lindsay, S. J.; Ning, Z. M.; Walter, K.; Zhang, Y. J.; Gerstein, M. B.; Abyzov, A.; Chen, J. M.; Clarke, D.; Lam, H.; Mu, X. J.; Sis, C.; Zhang, J.; Zhang, Y.; Gibbs, R. A.; Yu, F. L.; Bainbridge, M.; Challis, D.; Evani, U. S.; Kovar, C.; Lu, J.; Muzny, D.; Nagaswamy, U.; Reid, J. G.; Sabo, A.; Yu, J.; Guo, X. S.; Li, W. S.; Li, Y. R.; Wu, R. H.; Marth, G. T.; Garrison, E. P.; Leong, W. F.; Ward, A. N.; del Angel, G.; DePristo, M. A.; Gabriel, S. B.; Gupta, N.; Hartl, C.; Poplin, R. E.; Clark, A. G.; Rodriguez-Flores, J. L.; Flicek, P.; Clarke, L.; Smith, R. E.; Zheng-Bradley, X.; MacArthur, D. G.; Mardis, E. R.; Fulton, R.; Koboldt, D. C.; Gravel, S.; Bustamante, C. D.; Craig, D. W.; Christoforides, A.; Homer, N.; Izatt, T.; Sherry, S. T.; Xiao, C. L.; Dermitzakis, E. T.; Abecasis, G. R.; Kang, H. M.; McVean, G. A.; Gerstein, M. B.; Balasubramanian, S.; Habegger, L.; Yu, H. Y.; Flicek, P.; Clarke, L.; Cunningham, F.; Dunham, I.; Zerbino, D.; Zheng-Bradley, X.; Lage, K.; Jespersen, J. B.; Horn, H.; Montgomery, S. B.; DeGorter, M. K.; Khurana, E.; Tyler-Smith, C.; Chen, Y.; Colonna, V.; Xue, Y. L.; Gerstein, M. B.; Balasubramanian, S.; Fu, Y.; Kim, D.; Auton, A.; Marcketta, A.; Desalle, R.; Narechania, A.; Sayres, M. A. W.; Garrison, E. P.; Handsaker, R. E.; Kashin, S.; McCarroll, S. A.; Rodriguez-Flores, J. L.; Flicek, P.; Clarke, L.; Zheng-Bradley, X.; Erlich, Y.; Gymrek, M.; Willems, T. F.; Bustamante, C. D.; Mendez, F. L.; Poznik, G. D.; Underhill, P. A.; Lee, C.; Cerveira, E.; Malhotra, A.; Romanovitch, M.; Zhang, C. S.; Abecasis, G. R.; Coin, L.; Shao, H. J.; Mittelman, D.; Tyler-Smith, C.; Ayub, Q.; Banerjee, R.; Cerezo, M.; Chen, Y.; Fitzgerald, T.; Louzada, S.; Massaia, A.; McCarthy, S.; Ritchie, G. R.; Xue, Y. L.; Yang, F. T.; Gibbs, R. A.; Kovar, C.; Kalra, D.; Hale, W.; Muzny, D.; Reid, J. G.; Wang, J.; Dan, X.; Guo, X. S.; Li, G. Q.; Li, Y. R.; Ye, C.; Zheng, X. L.; Altshuler, D. M.; Flicek, P.; Clarke, L.; Zheng-Bradley, X.; Bentley, D. R.; Cox, A.; Humphray, S.; Kahn, S.; Sudbrak, R.; Albrecht, M. W.; Lienhard, M.; Larson, D.; Craig, D. W.; Izatt, T.; Kurdoglu, A. A.; Sherry, S. T.; Xiao, C. L.; Haussler, D.; Abecasis, G. R.; McVean, G. A.; Durbin, R. M.; Balasubramanian, S.; Keane, T. M.; McCarthy, S.; Stalker, J.; Chakravarti, A.; Knoppers, B. M.; Abecasis, G. R.; Barnes, K. C.; Beiswanger, C.; Burchard, E. G.; Bustamante, C. D.; Cai, H. Y.; Cao, H. Z.; Durbin, R. M.; Gerry, N. P.; Gharani, N.; Gibbs, R. A.; Gignoux, C. R.; Gravel, S.; Henn, B.; Jones, D.; Jorde, L.; Kaye, J. S.; Keinan, A.; Kent, A.; Kerasidou, A.; Li, Y. R.; Mathias, R.; McVean, G. A.; Moreno-Estrada, A.; Ossorio, P. N.; Parker, M.; Resch, A. M.; Rotimi, C. N.; Royal, C. D.; Sandoval, K.; Su, Y. Y.; Sudbrak, R.; Tian, Z. M.; Tishkoff, S.; Toji, L. H.; Tyler-Smith, C.; Via, M.; Wang, Y. H.; Yang, H. M.; Yang, L.; Zhu, J. Y.; Bodmer, W.; Bedoya, G.; Ruiz-Linares, A.; Cai, Z. M.; Gao, Y.; Chu, J. Y.; Peltonen, L.; Garcia-Montero, A.; Orfao, A.; Dutil, J.; Martinez-Cruzado, J. C.; Oleksyk, T. K.; Barnes, K. C.; Mathias, R. A.; Hennis, A.; Watson, H.; McKenzie, C.; Qadri, F.; LaRocque, R.; Sabeti, P. C.; Zhu, J. Y.;

- Deng, X. Y.; Sabeti, P. C.; Asogun, D.; Folarin, O.; Happi, C.; Omoniwa, O.; Stremlau, M.; Tariyal, R.; Jallow, M.; Joof, F. S.; Corrah, T.; Rockett, K.; Kwiatkowski, D.; Kooner, J.; Hien, T. T.; Dunstan, S. J.; Hang, N. T.; Fonniet, R.; Garry, R.; Kanneh, L.; Moses, L.; Sabeti, P. C.; Schieffelin, J.; Grant, D. S.; Gallo, C.; Poletti, G.; Saleheen, D.; Rasheed, A.; Brook, L. D.; Felsenfeld, A.; McEwen, J. E.; Vaydylevich, Y.; Green, E. D.; Duncanson, A.; Dunn, M.; Schloss, J. A.; Wang, J.; Yang, H. M.; Auton, A.; Brooks, L. D.; Durbin, R. M.; Garrison, E. P.; Kang, H. M.; Korbel, J. O.; Marchini, J. L.; McCarthy, S.; McVean, G. A.; Abecasis, G. R.; Consortium, G. P., A global reference for human genetic variation. *Nature* **2015**, *526* (7571), 68-+.
4. Genomes Project, C.; Auton, A.; Brooks, L. D.; Durbin, R. M.; Garrison, E. P.; Kang, H. M.; Korbel, J. O.; Marchini, J. L.; McCarthy, S.; McVean, G. A.; Abecasis, G. R., A global reference for human genetic variation. *Nature* **2015**, *526* (7571), 68-74.
5. Steinberg, M. H.; Sebastiani, P., Genetic modifiers of sickle cell disease. *Am J Hematol* **2012**, *87* (8), 795-803.
6. Bucossi, S.; Polimanti, R.; Mariani, S.; Ventriglia, M.; Bonvicini, C.; Migliore, S.; Manfellotto, D.; Salustri, C.; Vernieri, F.; Rossini, P. M.; Squitti, R., Association of K832R and R952K SNPs of Wilson's Disease Gene with Alzheimer's Disease. *J Alzheimers Dis* **2012**, *29* (4), 913-919.
7. Mistri, M.; Tamhankar, P. M.; Sheth, F.; Sanghavi, D.; Kondurkar, P.; Patil, S.; Idicula-Thomas, S.; Gupta, S.; Sheth, J., Identification of Novel Mutations in HEXA Gene in Children Affected with Tay Sachs Disease from India. *Plos One* **2012**, *7* (6).
8. Mills, R. E.; Luttig, C. T.; Larkins, C. E.; Beauchamp, A.; Tsui, C.; Pittard, W. S.; Devine, S. E., An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **2006**, *16* (9), 1182-1190.
9. Missirlis, P. I.; Mead, C. L.; Butland, S. L.; Ouellette, B. F.; Devon, R. S.; Leavitt, B. R.; Holt, R. A., Satellog: a database for the identification and prioritization of satellite repeats in disease association studies. *BMC Bioinformatics* **2005**, *6*, 145.
10. Vittori, A.; Breda, C.; Repici, M.; Orth, M.; Roos, R. A. C.; Outeiro, T. F.; Giorgini, F.; Hollox, E. J.; Hu, R. I. E., Copy-number variation of the neuronal glucose transporter gene SLC2A3 and age of onset in Huntington's disease. *Human Molecular Genetics* **2014**, *23* (12), 3129-3137.
11. Swaminathan, S.; Shen, L.; Kim, S.; Inlow, M.; West, J. D.; Faber, K. M.; Foroud, T.; Mayeux, R.; Saykin, A. J.; Alzheimer's Disease Neuroimaging, I.; Group, N.-L. N. F. S., Analysis of copy number variation in Alzheimer's disease: the NIALOAD/ NCRAD Family Study. *Curr Alzheimer Res* **2012**, *9* (7), 801-14.
12. Chung, B. H. Y.; Tao, V. Q.; Tso, W. W. Y., Copy number variation and autism: New insights and clinical implications. *J Formos Med Assoc* **2014**, *113* (7), 400-408.
13. Lakich, D.; Kazazian, H. H.; Antonarakis, S. E.; Gitschier, J., Inversions Disrupting the Factor-Viii Gene Are a Common-Cause of Severe Hemophilia-A. *Nat Genet* **1993**, *5* (3), 236-241.
14. Nickoloff, J. A.; De Haro, L. P.; Wray, J.; Hromas, R., Mechanisms of leukemia translocations. *Curr Opin Hematol* **2008**, *15* (4), 338-45.
15. Downing, J. R.; Head, D. R.; Parham, D. M.; Douglass, E. C.; Hulshof, M. G.; Link, M. P.; Motroni, T. A.; Grier, H. E.; Curcio-Brint, A. M.; Shapiro, D. N., Detection of the (11;22)(q24;q12) translocation of Ewing's sarcoma and peripheral neuroectodermal tumor by reverse transcription polymerase chain reaction. *Am J Pathol* **1993**, *143* (5), 1294-300.
16. Witte, J. S., Genome-Wide Association Studies and Beyond. *Annu Rev Publ Health* **2010**, *31*, 9-20.
17. Scott, M.; Gunderson, C. W.; Mateescu, E. M.; Zhang, Z. G.; Hwa, T., Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* **2010**, *330* (6007), 1099-1102.
18. Lee, C.; Roy, M., Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biology* **2004**, *5* (7).
19. Gardina, P. J.; Clark, T. A.; Shimada, B.; Staples, M. K.; Yang, Q.; Veitch, J.; Schweitzer, A.; Awad, T.; Sugnet, C.; Dee, S.; Davies, C.; Williams, A.; Turpaz, Y., Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *Bmc Genomics* **2006**, *7*.
20. Holloway, B.; Luck, S.; Beatty, M.; Rafalski, J. A.; Li, B. L., Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *Bmc Genomics* **2011**, *12*.
21. Tarca, A. L.; Romero, R.; Draghici, S., Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol* **2006**, *195* (2), 373-388.
22. Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O., Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. *Science* **1995**, *270* (5235), 467-470.
23. Nagalakshmi, U.; Wang, Z.; Waern, K.; Shou, C.; Raha, D.; Gerstein, M.; Snyder, M., The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **2008**, *320* (5881), 1344-1349.
24. Sui, Y. X.; Zhao, X. Y.; Speed, T. P.; Wu, Z. J., Background Adjustment for DNA Microarrays Using a Database of Microarray Experiments. *J Comput Biol* **2009**, *16* (11), 1501-1515.

25. Haas, B. J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P. D.; Bowden, J.; Couger, M. B.; Eccles, D.; Li, B.; Lieber, M.; MacManes, M. D.; Ott, M.; Orvis, J.; Pochet, N.; Strozzi, F.; Weeks, N.; Westerman, R.; William, T.; Dewey, C. N.; Henschel, R.; Leduc, R. D.; Friedman, N.; Regev, A., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **2013**, 8 (8), 1494-1512.
26. Anders, S.; Huber, W., Differential expression analysis for sequence count data. *Genome Biol* **2010**, 11 (10), R106.
27. Robinson, M. D.; Oshlack, A., A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **2010**, 11 (3), R25.
28. Love, M. I.; Huber, W.; Anders, S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **2014**, 15 (12), 550.
29. Hardcastle, T. J.; Kelly, K. A., baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *Bmc Bioinformatics* **2010**, 11.
30. Leng, N.; Dawson, J. A.; Thomson, J. A.; Ruotti, V.; Rissman, A. I.; Smits, B. M.; Haag, J. D.; Gould, M. N.; Stewart, R. M.; Kendzierski, C., EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **2013**, 29 (8), 1035-43.
31. Law, C. W.; Chen, Y.; Shi, W.; Smyth, G. K., voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **2014**, 15 (2), R29.
32. Tarazona, S.; Furio-Tari, P.; Turra, D.; Pietro, A. D.; Nueda, M. J.; Ferrer, A.; Conesa, A., Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* **2015**, 43 (21), e140.
33. Li, J.; Tibshirani, R., Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* **2013**, 22 (5), 519-36.
34. Boroughs, L. K.; DeBerardinis, R. J., Metabolic pathways promoting cancer cell survival and growth. *Nat Cell Biol* **2015**, 17 (4), 351-9.
35. Ein-Dor, L.; Kela, I.; Getz, G.; Givol, D.; Domany, E., Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **2005**, 21 (2), 171-178.
36. Wu, G.; Zhi, D., Pathway-based approaches for sequencing-based genome-wide association studies. *Genet Epidemiol* **2013**, 37 (5), 478-94.
37. Bader, G. D.; Cary, M. P.; Sander, C., Pathguide: a Pathway Resource List. *Nucleic Acids Research* **2006**, 34, D504-D506.
38. Huang, D. W.; Sherman, B. T.; Tan, Q.; Collins, J. R.; Alvord, W. G.; Roayaei, J.; Stephens, R.; Baseler, M. W.; Lane, H. C.; Lempicki, R. A., The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* **2007**, 8 (9), R183.
39. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **2005**, 102 (43), 15545-50.
40. Carbon, S.; Dietze, H.; Lewis, S. E.; Mungall, C. J.; Munoz-Torres, M. C.; Basu, S.; Chisholm, R. L.; Dodson, R. J.; Fey, P.; Thomas, P. D.; Mi, H.; Muruganujan, A.; Huang, X.; Poudel, S.; Hu, J. C.; Aleksander, S. A.; McIntosh, B. K.; Renfro, D. P.; Siegle, D. A.; Antonazzo, G.; Attrill, H.; Brown, N. H.; Marygold, S. J.; McQuilton, P.; Ponting, L.; Millburn, G. H.; Rey, A. J.; Stefancsik, R.; Tweedie, S.; Falls, K.; Schroeder, A. J.; Courtot, M.; Osumi-Sutherland, D.; Parkinson, H.; Roncaglia, P.; Lovering, R. C.; Foulger, R. E.; Huntley, R. P.; Denny, P.; Campbell, N. H.; Kramarz, B.; Patel, S.; Buxton, J. L.; Umrao, Z.; Deng, A. T.; Alrohaif, H.; Mitchell, K.; Ratnaraj, F.; Omer, W.; Rodriguez-Lopez, M.; Chibucos, M. C.; Giglio, M.; Nadendla, S.; Duesbury, M. J.; Koch, M.; Meldal, B. H. M.; Melidoni, A.; Porras, P.; Orchard, S.; Shrivastava, A.; Chang, H. Y.; Finn, R. D.; Fraser, M.; Mitchell, A. L.; Nuka, G.; Potter, S.; Rawlings, N. D.; Richardson, L.; Sangrador-Vegas, A.; Young, S. Y.; Blake, J. A.; Christie, K. R.; Dolan, M. E.; Drabkin, H. J.; Hill, D. P.; Ni, L.; Sitnikov, D.; Harris, M. A.; Hayles, J.; Oliver, S. G.; Rutherford, K.; Wood, V.; Bahler, J.; Lock, A.; De Pons, J.; Dwinell, M.; Shimoyama, M.; Lalederkind, S.; Hayman, G. T.; Tutaj, M.; Wang, S. J.; D'Eustachio, P.; Matthews, L.; Balhoff, J. P.; Balakrishnan, R.; Binkley, G.; Cherry, J. M.; Costanzo, M. C.; Engel, S. R.; Miyasato, S. R.; Nash, R. S.; Simison, M.; Skrzypek, M. S.; Weng, S.; Wong, E. D.; Feuermann, M.; Gaudet, P.; Berardini, T. Z.; Li, D.; Muller, B.; Reiser, L.; Huala, E.; Argasinska, J.; Arighi, C.; Auchincloss, A.; Axelsen, K.; Argoud-Puy, G.; Bateman, A.; Bely, B.; Blatter, M. C.; Bonilla, C.; Bougueleret, L.; Boutet, E.; Breuza, L.; Bridge, A.; Britto, R.; Hye-A-Bye, H.; Casals, C.; Cibrian-Uhalte, E.; Coudert, E.; Cusin, I.; Duek-Roggli, P.; Estreicher, A.; Famiglietti, L.; Gane, P.; Garmiri, P.; Georghiou, G.; Gos, A.; Gruaz-Gumowski, N.; Hatton-Ellis, E.; Hinz, U.; Holmes, A.; Hulo, C.; Jungo, F.; Keller, G.; Laiho, K.; Lemercier, P.; Lieberherr, D.; MacDougall, A.; Magrane, M.; Martin, M. J.; Masson, P.; Natale, D. A.; O'Donovan, V.; Pedruzzi, I.; Pichler, K.; Poggioli, D.; Poux, S.; Rivoire, C.; Roechert, B.; Sawford, T.; Schneider, M.; Speretta, E.; Shypitsyna, A.; Stutz, A.; Sundaram, S.; Tognolli, M.; Wu, C.;

- Xenarios, I.; Yeh, L. S.; Chan, J.; Gao, S.; Howe, K.; Kishore, R.; Lee, R.; Li, Y.; Lomax, J.; Muller, H. M.; Raciti, D.; Van Auken, K.; Berriman, M.; Stein, L.; Kersey, P.; Sternberg, P. W.; Howe, D.; Westerfield, M.; Consortium, G. O., Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **2017**, *45* (D1), D331-D338.
41. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K., KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **2017**, *45* (D1), D353-D361.
42. Fabregat, A.; Sidiropoulos, K.; Viteri, G.; Forner, O.; Marin-Garcia, P.; Arnau, V.; D'Eustachio, P.; Stein, L.; Hermjakob, H., Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **2017**, *18* (1), 142.
43. Gama-Castro, S.; Salgado, H.; Santos-Zavaleta, A.; Ledezma-Tejeida, D.; Muniz-Rascado, L.; Garcia-Sotelo, J. S.; Alquicira-Hernandez, K.; Martinez-Flores, I.; Pannier, L.; Castro-Mondragon, J. A.; Medina-Rivera, A.; Solano-Lira, H.; Bonavides-Martinez, C.; Perez-Rueda, E.; Alquicira-Hernandez, S.; Porron-Sotelo, L.; Lopez-Fuentes, A.; Hernandez-Koutoucheva, A.; Del Moral-Chavez, V.; Rinaldi, F.; Collado-Vides, J., RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research* **2016**, *44* (D1), D133-D143.
44. Mi, H.; Thomas, P., PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol* **2009**, *563*, 123-40.
45. Kramer, A.; Green, J.; Pollard, J., Jr.; Tugendreich, S., Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2014**, *30* (4), 523-30.
46. Anthony, K.; Skinner, M. A.; Buchoff, J. R.; McCarthy, N.; Schaefer, C. F.; Buetow, K. H., The NCI-Nature Pathway Interaction Database: A comprehensive resource for cell signaling information. *Cancer Research* **2011**, *71*.
47. Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Melius, J.; Cirillo, E.; Coort, S. L.; Digles, D.; Ehrhart, F.; Giesbertz, P.; Kalafati, M.; Martens, M.; Miller, R.; Nishida, K.; Rieswijk, L.; Waagmeester, A.; Eijssen, L. M. T.; Evelo, C. T.; Pico, A. R.; Willighagen, E. L., WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **2018**, *46* (D1), D661-D667.
48. Jewison, T.; Su, Y. L.; Disfany, F. M.; Liang, Y. J.; Knox, C.; Maciejewski, A.; Poelzer, J.; Huynh, J.; Zhou, Y.; Arndt, D.; Djoumbou, Y.; Liu, Y. F.; Deng, L.; Guo, A. C.; Han, B.; Pon, A.; Wilson, M.; Rafatnia, S.; Liu, P.; Wishart, D. S., SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. *Nucleic Acids Research* **2014**, *42* (D1), D478-D484.
49. Kamburov, A.; Stelzl, U.; Lehrach, H.; Herwig, R., The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research* **2013**, *41* (D1), D793-D800.
50. Cerami, E. G.; Gross, B. E.; Demir, E.; Rodchenkov, I.; Babur, O.; Anwar, N.; Schultz, N.; Bader, G. D.; Sander, C., Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* **2011**, *39*, D685-D690.
51. Tarca, A. L.; Draghici, S.; Khatr, P.; Hassan, S. S.; Mittal, P.; Kim, J. S.; Kim, C. J.; Kusanovic, J. P.; Romero, R., A novel signaling pathway impact analysis. *Bioinformatics* **2009**, *25* (1), 75-82.
52. Gu, Z.; Liu, J.; Cao, K.; Zhang, J.; Wang, J., Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Syst Biol* **2012**, *6*, 56.
53. Dutta, B.; Wallqvist, A.; Reifman, J., PathNet: a tool for pathway analysis using topological information. *Source Code Biol Med* **2012**, *7* (1), 10.
54. Bayerlova, M.; Jung, K.; Kramer, F.; Klemm, F.; Bleckmann, A.; Beissbarth, T., Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics* **2015**, *16*, 334.
55. Pavlidis, P.; Qin, J.; Arango, V.; Mann, J. J.; Sibille, E., Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* **2004**, *29* (6), 1213-1222.
56. Tian, L.; Greenberg, S. A.; Kong, S. W.; Altschuler, J.; Kohane, I. S.; Park, P. J., Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA* **2005**, *102* (38), 13544-9.
57. Kim, S. Y.; Volsky, D. J., PAGE: Parametric analysis of gene set enrichment. *Bmc Bioinformatics* **2005**, *6*.
58. Jiang, Z.; Gentleman, R., Extensions to gene set enrichment. *Bioinformatics* **2007**, *23* (3), 306-13.
59. Efron, B.; Tibshirani, R., On Testing the Significance of Sets of Genes. *Ann Appl Stat* **2007**, *1* (1), 107-129.
60. Goeman, J. J.; van de Geer, S. A.; de Kort, F.; van Houwelingen, H. C., A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **2004**, *20* (1), 93-99.

61. Nam, D.; Kim, J.; Kim, S. Y.; Kim, S., GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Research* **2010**, *38*, W749-W754.
62. Segre, A. V.; Consortium, D.; investigators, M.; Groop, L.; Mootha, V. K.; Daly, M. J.; Altshuler, D., Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *Plos Genet* **2010**, *6* (8).
63. Lee, P. H.; O'Dushlaine, C.; Thomas, B.; Purcell, S. M., INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **2012**, *28* (13), 1797-1799.
64. Kofler, R.; Schlotterer, C., Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* **2012**, *28* (15), 2084-2085.
65. de Leeuw, C. A.; Mooij, J. M.; Heskes, T.; Posthuma, D., MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **2015**, *11* (4), e1004219.
66. Zhang, K. L.; Chang, S. H.; Guo, L. Y.; Wang, J., I-GSEA4GWAS v2: a web server for functional analysis of SNPs in trait-associated pathways identified from genome-wide association study. *Protein Cell* **2015**, *6* (3), 221-224.
67. Yoon, S.; Nguyen, H. C. T.; Yoo, Y. J.; Kim, J.; Baik, B.; Kim, S.; Kim, J.; Kim, S.; Nam, D., Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Research* **2018**, *46* (10).
68. Zhang, H.; Wheeler, W.; Hyland, P. L.; Yang, Y.; Shi, J.; Chatterjee, N.; Yu, K., A Powerful Procedure for Pathway-Based Meta-analysis Using Summary Statistics Identifies 43 Pathways Associated with Type II Diabetes in European Populations. *Plos Genet* **2016**, *12* (6), e1006122.
69. Xenarios, I.; Salwinski, L.; Duan, X. Q. J.; Higney, P.; Kim, S. M.; Eisenberg, D., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **2002**, *30* (1), 303-305.
70. Bader, G. D.; Betel, D.; Hogue, C. W. V., BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research* **2003**, *31* (1), 248-250.
71. Stark, C.; Breitkreutz, B. J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M., BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **2006**, *34*, D535-D539.
72. Peri, S.; Navarro, J. D.; Kristiansen, T. Z.; Amanchy, R.; Surendranath, V.; Muthusamy, B.; Gandhi, T. K. B.; Chandrika, K. N.; Deshpande, N.; Suresh, S.; Rashmi, B. P.; Shanker, K.; Padma, N.; Niranjana, V.; Harsha, H. C.; Talreja, N.; Vrushabendra, B. M.; Ramya, M. A.; Yatish, A. J.; Joy, M.; Shivashankar, H. N.; Kavitha, M. P.; Menezes, M.; Choudhury, D. R.; Ghosh, N.; Saravana, R.; Chandran, S.; Mohan, S.; Jonnalagadda, C. K.; Prasad, C. K.; Kumar-Sinha, C.; Deshpande, K. S.; Pandey, A., Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research* **2004**, *32*, D497-D501.
73. Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardoza, A. P.; Santonico, E.; Castagnoli, L.; Cesareni, G., MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research* **2012**, *40* (D1), D857-D861.
74. Pagel, P.; Kovac, S.; Oesterheld, M.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Mark, P.; Stumpflen, V.; Mewes, H. W.; Ruepp, A.; Frishman, D., The MIPS mammalian protein-protein interaction database. *Bioinformatics* **2005**, *21* (6), 832-834.
75. Alanis-Lobato, G.; Andrade-Navarro, M. A.; Schaefer, M. H., HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res* **2017**, *45* (D1), D408-D414.
76. Szklarczyk, D.; Morris, J. H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N. T.; Roth, A.; Bork, P.; Jensen, L. J.; von Mering, C., The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research* **2017**, *45* (D1), D362-D368.
77. Karp, P. D.; Riley, M.; Paley, S. M.; PelligriniToole, A., EcoCyc: An encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Research* **1996**, *24* (1), 32-39.
78. Whitaker, J. W.; Letunic, I.; McConkey, G. A.; Westhead, D. R., metaTIGER: a metabolic evolution resource. *Nucleic Acids Research* **2009**, *37*, D531-D538.
79. Wang, X.; Cairns, M. J., SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. *Bioinformatics* **2014**, *30* (12), 1777-1779.
80. Yoon, S.; Kim, S. Y.; Nam, D., Improving Gene-Set Enrichment Analysis of RNA-Seq Data with Small Replicates. *PLoS One* **2016**, *11* (11), e0165919.
81. Wang, Z.; Gerstein, M.; Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **2009**, *10* (1), 57-63.
82. Marioni, J. C.; Mason, C. E.; Mane, S. M.; Stephens, M.; Gilad, Y., RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **2008**, *18* (9), 1509-17.
83. Dillies, M. A.; Rau, A.; Aubert, J.; Hennequet-Antier, C.; Jeanmougin, M.; Servant, N.; Keime, C.; Marot, G.; Castel, D.; Estelle, J.; Guernec, G.; Jagla, B.; Jouneau, L.; Laloe, D.; Le Gall, C.; Schaeffer, B.; Le

- Crom, S.; Guedj, M.; Jaffrezic, F.; French StatOmique, C., A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **2013**, *14* (6), 671-83.
84. Robinson, M. D.; Smyth, G. K., Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **2007**, *23* (21), 2881-7.
 85. Anders, S.; Huber, W., Differential expression analysis for sequence count data. *Genome Biology* **2010**, *11* (10).
 86. Bullard, J. H.; Purdom, E.; Hansen, K. D.; Dudoit, S., Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **2010**, *11*, 94.
 87. Rapaport, F.; Khanin, R.; Liang, Y.; Pirun, M.; Krek, A.; Zumbo, P.; Mason, C. E.; Succi, N. D.; Betel, D., Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **2013**, *14* (9), R95.
 88. Zeeberg, B. R.; Feng, W.; Wang, G.; Wang, M. D.; Fojo, A. T.; Sunshine, M.; Narasimhan, S.; Kane, D. W.; Reinhold, W. C.; Lababidi, S.; Bussey, K. J.; Riss, J.; Barrett, J. C.; Weinstein, J. N., GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **2003**, *4* (4), R28.
 89. Huang da, W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **2009**, *37* (1), 1-13.
 90. Carver, B. S.; Chapinski, C.; Wongvipat, J.; Hieronymus, H.; Chen, Y.; Chandralapaty, S.; Arora, V. K.; Le, C.; Koutcher, J.; Scher, H.; Scardino, P. T.; Rosen, N.; Sawyers, C. L., Reciprocal feedback regulation of PI3K and androgen receptor signaling in PTEN-deficient prostate cancer. *Cancer Cell* **2011**, *19* (5), 575-86.
 91. Schwarz, J. K.; Payton, J. E.; Rashmi, R.; Xiang, T.; Jia, Y. H.; Huettner, P.; Rogers, B. E.; Yang, Q.; Watson, M.; Rader, J. S.; Grigsby, P. W., Pathway-Specific Analysis of Gene Expression Data Identifies the PI3K/Akt Pathway as a Novel Therapeutic Target in Cervical Cancer. *Clinical Cancer Research* **2012**, *18* (5), 1464-1471.
 92. Li, H. L.; Chiappinelli, K. B.; Guzzetta, A. A.; Easwaran, H.; Yen, R. W. C.; Vatapalli, R.; Topper, M. J.; Luo, J. J.; Connolly, R. M.; Azad, N. S.; Stearns, V.; Pardoll, D. M.; Davidson, N.; Jones, P. A.; Slamon, D. J.; Baylin, S. B.; Zahnow, C. A.; Ahuja, N., Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget* **2014**, *5* (3), 587-598.
 93. Wang, X.; Cairns, M. J., Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *Bmc Bioinformatics* **2013**, *14*.
 94. Xiong, Q.; Mukherjee, S.; Furey, T. S., GSASeqSP: A Toolset for Gene Set Association Analysis of RNA-Seq Data. *Scientific Reports* **2014**, *4*.
 95. Lee, C.; Patil, S.; Sartor, M. A., RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power. *Bioinformatics* **2015**.
 96. Feng, J. X.; Meyer, C. A.; Wang, Q.; Liu, J. S.; Liu, X. S.; Zhang, Y., GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* **2012**, *28* (21), 2782-2788.
 97. Nam, D., Effect of the absolute statistic on gene-sampling gene-set analysis methods. *Stat Methods Med Res* **2015**.
 98. Robinson, M. D.; McCarthy, D. J.; Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26* (1), 139-140.
 99. Goeman, J. J.; Buhlmann, P., Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **2007**, *23* (8), 980-987.
 100. Newton, M. A.; Quintana, F. A.; Den Boon, J. A.; Sengupta, S.; Ahlquist, P., Random-Set Methods Identify Distinct Aspects of the Enrichment Signal in Gene-Set Analysis. *Ann Appl Stat* **2007**, *1* (1), 85-106.
 101. Wu, D.; Smyth, G. K., Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research* **2012**, *40* (17).
 102. Nam, D., De-correlating expression in gene-set analysis. *Bioinformatics* **2010**, *26* (18), i511-i516.
 103. Nam, D.; Kim, S. Y., Gene-set approach for expression pattern analysis (vol 9, pg 189, 2008). *Briefings in Bioinformatics* **2008**, *9* (5), 450-450.
 104. Saxena, V.; Orgill, D.; Kohane, I., Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Research* **2006**, *34* (22).
 105. Smyth, G. K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **2004**, *3* (1), 1-25.
 106. Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K., limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **2015**, *43* (7), e47.
 107. Song, S.; Black, M. A., Microarray-based gene set analysis: a comparison of current methods. *Bmc Bioinformatics* **2008**, *9*.
 108. Cancer Genome Atlas Research, N., Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **2013**, *499* (7456), 43-9.

109. Kovalchik, S., RISmed: Download Content from NCBI Databases. **2015**.
110. R.Core.Team, R: A Language and Environment for Statistical Computing. 2015.
111. Eddebuettel, D.; Francois, R., Rcpp: Seamless R and C plus plus Integration. *J Stat Softw* **2011**, *40* (8), 1-18.
112. Pickrell, J. K.; Marioni, J. C.; Pai, A. A.; Degner, J. F.; Engelhardt, B. E.; Nkadori, E.; Veyrieras, J. B.; Stephens, M.; Gilad, Y.; Pritchard, J. K., Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **2010**, *464* (7289), 768-772.
113. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdottir, H.; Tamayo, P.; Mesirov, J. P., Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27* (12), 1739-40.
114. Gray, K. A.; Yates, B.; Seal, R. L.; Wright, M. W.; Bruford, E. A., Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research* **2015**, *43* (D1), D1079-D1085.
115. Jiang, H.; Bivens, N. J.; Ries, J. E.; Whitworth, K. M.; Green, J. A.; Forrester, L. J.; Springer, G. K.; Didion, B. A.; Mathialagan, N.; Prather, R. S.; Lucy, M. C., Constructing cDNA libraries with fewer clones that contain long poly(dA) tails. *Biotechniques* **2001**, *31* (1), 38-+.
116. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D. R.; Pimentel, H.; Salzberg, S. L.; Rinn, J. L.; Pachter, L., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (vol 7, pg 562, 2012). *Nat Protoc* **2014**, *9* (10), 2513-2513.
117. Barry, W. T.; Nobel, A. B.; Wright, F. A., A Statistical Framework for Testing Functional Categories in Microarray Data. *Ann Appl Stat* **2008**, *2* (1), 286-315.
118. Koboldt, D. C.; Fulton, R. S.; McLellan, M. D.; Schmidt, H.; Kalicki-Veizer, J.; McMichael, J. F.; Fulton, L. L.; Dooling, D. J.; Ding, L.; Mardis, E. R.; Wilson, R. K.; Ally, A.; Balasundaram, M.; Butterfield, Y. S. N.; Carlsen, R.; Carter, C.; Chu, A.; Chuah, E.; Chun, H. J. E.; Coope, R. J. N.; Dhalla, N.; Guin, R.; Hirst, C.; Hirst, M.; Holt, R. A.; Lee, D.; Li, H. Y. I.; Mayo, M.; Moore, R. A.; Mungall, A. J.; Pleasance, E.; Robertson, A. G.; Schein, J. E.; Shafiei, A.; Sipahimalani, P.; Slobodan, J. R.; Stoll, D.; Tam, A.; Thiessen, N.; Varhol, R. J.; Wye, N.; Zeng, T.; Zhao, Y. J.; Birol, I.; Jones, S. J. M.; Marra, M. A.; Cherniack, A. D.; Saksena, G.; Onofrio, R. C.; Pho, N. H.; Carter, S. L.; Schumacher, S. E.; Tabak, B.; Hernandez, B.; Gentry, J.; Nguyen, H.; Crenshaw, A.; Ardlie, K.; Beroukhi, R.; Winckler, W.; Getz, G.; Gabriel, S. B.; Meyerson, M.; Chin, L.; Park, P. J.; Kucherlapati, R.; Hoadley, K. A.; Auman, J. T.; Fan, C.; Turman, Y. J.; Shi, Y.; Li, L.; Topal, M. D.; He, X. P.; Chao, H. H.; Prat, A.; Silva, G. O.; Iglesia, M. D.; Zhao, W.; Usary, J.; Berg, J. S.; Adams, M.; Booker, J.; Wu, J. Y.; Gulabani, A.; Bodenheimer, T.; Hoyle, A. P.; Simons, J. V.; Soloway, M. G.; Mose, L. E.; Jefferys, S. R.; Balu, S.; Parker, J. S.; Hayes, D. N.; Perou, C. M.; Malik, S.; Mahurkar, S.; Shen, H.; Weisenberger, D. J.; Triche, T.; Lai, P. H.; Bootwalla, M. S.; Maglinte, D. T.; Berman, B. P.; Van den Berg, D. J.; Baylin, S. B.; Laird, P. W.; Creighton, C. J.; Donehower, L. A.; Getz, G.; Noble, M.; Voet, D.; Saksena, G.; Gehlenborg, N.; DiCara, D.; Zhang, J. H.; Zhang, H. L.; Wu, C. J.; Liu, S. Y.; Lawrence, M. S.; Zou, L. H.; Sivachenko, A.; Lin, P.; Stojanov, P.; Jing, R.; Cho, J.; Sinha, R.; Park, R. W.; Nazaire, M. D.; Robinson, J.; Thorvaldsdottir, H.; Mesirov, J.; Park, P. J.; Chin, L.; Reynolds, S.; Kreisberg, R. B.; Bernard, B.; Bressler, R.; Erkkila, T.; Lin, J.; Thorsson, V.; Zhang, W.; Shmulevich, I.; Ciriello, G.; Weinhold, N.; Schultz, N.; Gao, J. J.; Cerami, E.; Gross, B.; Jacobsen, A.; Sinha, R.; Aksoy, B. A.; Antipin, Y.; Reva, B.; Shen, R. L.; Taylor, B. S.; Ladanyi, M.; Sander, C.; Anur, P.; Spellman, P. T.; Lu, Y. L.; Liu, W. B.; Verhaak, R. R. G.; Mills, G. B.; Akbani, R.; Zhang, N. X.; Broom, B. M.; Casasent, T. D.; Wakefield, C.; Unruh, A. K.; Baggerly, K.; Coombes, K.; Weinstein, J. N.; Haussler, D.; Benz, C. C.; Stuart, J. M.; Benz, S. C.; Zhu, J. C.; Szeto, C. C.; Scott, G. K.; Yau, C.; Paul, E. O.; Carlin, D.; Wong, C.; Sokolov, A.; Thüsenberg, J.; Mooney, S.; Ng, S.; Goldstein, T. C.; Ellrott, K.; Grifford, M.; Wilks, C.; Ma, S.; Craft, B.; Yan, C. H.; Hu, Y.; Meerzaman, D.; Gastier-Foster, J. M.; Bowen, J.; Ramirez, N. C.; Black, A. D.; Pyatt, R. E.; White, P.; Zmuda, E. J.; Frick, J.; Lichtenberg, T.; Brookens, R.; George, M. M.; Gerken, M. A.; Harper, H. A.; Leraas, K. M.; Wise, L. J.; Tabler, T. R.; McAllister, C.; Barr, T.; Hart-Kothari, M.; Tarvin, K.; Saller, C.; Sandusky, G.; Mitchell, C.; Iacocca, M. V.; Brown, J.; Rabeno, B.; Czerwinski, C.; Petrelli, N.; Dolzhansky, O.; Abramov, M.; Voronina, O.; Potapova, O.; Marks, J. R.; Suchorska, W. M.; Murawa, D.; Kyler, W.; Ibbs, M.; Korski, K.; Sychala, A.; Murawa, P.; Brzezinski, J. J.; Perz, H.; Lazniak, R.; Teresiak, M.; Tatka, H.; Leporowska, E.; Bogusz-Czerniewicz, M.; Malicki, J.; Mackiewicz, A.; Wiznerowicz, M.; Le, X. V.; Kohl, B.; Tien, N. V.; Thorp, R.; Bang, N. V.; Sussman, H.; Phu, B. D.; Hajek, R.; Hung, N. P.; Tran, V. T. P.; Thang, H. Q.; Khan, K. Z.; Penny, R.; Mallery, D.; Curley, E.; Shelton, C.; Yena, P.; Ingle, J. N.; Couch, F. J.; Lingle, W. L.; King, T. A.; Gonzalez-Angulo, A. M.; Mills, G. B.; Dyer, M. D.; Liu, S. Y.; Meng, X. L.; Patangan, M.; Waldman, F.; Stoppler, H.; Rathmell, W. K.; Thorne, L.; Huang, M.; Boice, L.; Hill, A.; Morrison, C.; Gaudioso, C.; Bshara, W.; Daily, K.; Egea, S. C.; Pegram, M. D.; Gomez-Fernandez, C.; Dhir, R.; Bhargava, R.; Brufsky, A.; Shriver, C. D.; Hooke, J. A.; Campbell, J. L.; Mural, R. J.; Hu, H.; Somiari, S.; Larson, C.; Deyarmin, B.; Kvecher, L.; Kovatich, A. J.; Ellis, M. J.; King, T. A.; Hu, H.; Couch, F. J.; Mural, R. J.; Stricker, T.; White, K.; Olopade, O.; Ingle, J. N.; Luo, C. Q.; Chen, Y. Q.; Marks, J. R.; Waldman, F.; Wiznerowicz, M.; Bose, R.; Chang, L. W.; Beck, A. H.; Gonzalez-Angulo, A. M.; Pihl, T.; Jensen, M.; Sfeir, R.; Kahn, A.; Chu, A.; Kothiyal, P.; Wang, Z. N.; Snyder, E.; Pontius,

J.; Ayala, B.; Backus, M.; Walton, J.; Baboud, J.; Berton, D.; Nicholls, M.; Srinivasan, D.; Raman, R.; Girshik, S.; Kigonya, P.; Alonso, S.; Sanbhadti, R.; Barletta, S.; Pot, D.; Sheth, M.; Demchok, J. A.; Shaw, K. R. M.; Yang, L. M.; Eley, G.; Ferguson, M. L.; Tarnuzzer, R. W.; Zhang, J. S.; Dillon, L. A. L.; Buetow, K.; Fielding, P.; Ozenberger, B. A.; Guyer, M. S.; Sofia, H. J.; Palchik, J. D.; Network, C. G. A., Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490* (7418), 61-70.

119. Young, M. D.; Wakefield, M. J.; Smyth, G. K.; Oshlack, A., Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **2010**, *11* (2).

120. Li, H.; Lovci, M. T.; Kwon, Y. S.; Rosenfeld, M. G.; Fu, X. D.; Yeo, G. W., Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *P Natl Acad Sci USA* **2008**, *105* (51), 20179-20184.

121. Yano, K.; Yamamoto, E.; Aya, K.; Takeuchi, H.; Lo, P. C.; Hu, L.; Yamasaki, M.; Yoshida, S.; Kitano, H.; Hirano, K.; Matsuoka, M., Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature genetics* **2016**, *48* (8), 927-34.

122. Wang, K.; Li, M.; Hakonarson, H., Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* **2010**, *11* (12), 843-54.

123. Liu, J. Z.; McRae, A. F.; Nyholt, D. R.; Medland, S. E.; Wray, N. R.; Brown, K. M.; Hayward, N. K.; Montgomery, G. W.; Visscher, P. M.; Martin, N. G.; Macgregor, S., A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* **2010**, *87* (1), 139-45.

124. Li, M. X.; Gui, H. S.; Kwan, J. S.; Sham, P. C., GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet* **2011**, *88* (3), 283-93.

125. Wang, K.; Li, M.; Bucan, M., Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **2007**, *81* (6), 1278-83.

126. Holden, M.; Deng, S.; Wojnowski, L.; Kulle, B., GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics (Oxford, England)* **2008**, *24* (23), 2784-5.

127. Zhang, K.; Cui, S.; Chang, S.; Zhang, L.; Wang, J., i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic acids research* **2010**, *38* (Web Server issue), W90-5.

128. Weng, L.; Macciardi, F.; Subramanian, A.; Guffanti, G.; Potkin, S. G.; Yu, Z.; Xie, X., SNP-based pathway enrichment analysis for genome-wide association studies. *BMC bioinformatics* **2011**, *12*, 99.

129. Holmans, P.; Green, E. K.; Pahwa, J. S.; Ferreira, M. A.; Purcell, S. M.; Sklar, P.; Owen, M. J.; O'Donovan, M. C.; Craddock, N., Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* **2009**, *85* (1), 13-24.

130. Kwak, I. Y.; Pan, W., Adaptive gene- and pathway-trait association testing with GWAS summary statistics. *Bioinformatics (Oxford, England)* **2016**, *32* (8), 1178-84.

131. Wang, Q.; Yu, H.; Zhao, Z.; Jia, P., EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics (Oxford, England)* **2015**, *31* (15), 2591-4.

132. Rossin, E. J.; Lage, K.; Raychaudhuri, S.; Xavier, R. J.; Tatar, D.; Benita, Y.; Cotsapas, C.; Daly, M. J., Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS genetics* **2011**, *7* (1), e1001273.

133. Haw, R.; Hermjakob, H.; D'Eustachio, P.; Stein, L., Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **2011**, *11* (18), 3598-613.

134. Morris, A. P.; Voight, B. F.; Teslovich, T. M.; Ferreira, T.; Segre, A. V.; Steinthorsdottir, V.; Strawbridge, R. J.; Khan, H.; Grallert, H.; Mahajan, A.; Prokopenko, I.; Kang, H. M.; Dina, C.; Esko, T.; Fraser, R. M.; Kanoni, S.; Kumar, A.; Lagou, V.; Langenberg, C.; Luan, J. A.; Lindgren, C. M.; Muller-Nurasyid, M.; Pechlivanis, S.; Rayner, N. W.; Scott, L. J.; Wiltshire, S.; Yengo, L.; Kinnunen, L.; Rossin, E. J.; Raychaudhuri, S.; Johnson, A. D.; Dimas, A. S.; Loos, R. J. F.; Vedantam, S.; Chen, H.; Florez, J. C.; Fox, C.; Liu, C. T.; Rybin, D.; Couper, D. J.; Kao, W. H. L.; Li, M.; Cornelis, M. C.; Kraft, P.; Sun, Q.; van Dam, R. M.; Stringham, H. M.; Chines, P. S.; Fischer, K.; Fontanillas, P.; Holmen, O. L.; Hunt, S. E.; Jackson, A. U.; Kong, A.; Lawrence, R.; Meyer, J.; Perry, J. R. B.; Platou, C. G. P.; Potter, S.; Rehnberg, E.; Robertson, N.; Sivapalaratnam, S.; Stancakova, A.; Stirrups, K.; Thorleifsson, G.; Tikkanen, E.; Wood, A. R.; Almgren, P.; Atalay, M.; Benediktsson, R.; Bonnycastle, L. L.; Burt, N.; Carey, J.; Charpentier, G.; Crenshaw, A. T.; Doney, A. S. F.; Dorkhan, M.; Edkins, S.; Emilsson, V.; Eury, E.; Forsen, T.; Gertow, K.; Gigante, B.; Grant, G. B.; Groves, C. J.; Guiducci, C.; Herder, C.; Hreidarsson, A. B.; Hui, J. N.; James, A.; Jonsson, A.; Rathmann, W.; Klopp, N.; Kravic, J.; Krjutskov, K.; Langford, C.; Leander, K.; Lindholm, E.; Lobbens, S.; Mannisto, S.; Mirza, G.; Muhleisen, T. W.; Musk, B.; Parkin, M.; Rallidis, L.; Saramies, J.; Sennblad, B.; Shah, S.; Sigurdsson, G.; Silveira, A.; Steinbach, G.; Thorand, B.; Trakalo, J.; Veglia, F.; Wennauer, R.; Winckler, W.; Zabaneh, D.; Campbell, H.; van Duijn, C.; Uitterlinden, A. G.; Hofman, A.; Sijbrands, E.; Abecasis, G. R.; Owen, K. R.; Zeggini, E.; Trip, M. D.; Forouhi, N. G.; Syvanen, A. C.; Eriksson, J. G.; Peltonen, L.; Nothen, M. M.; Balkau, B.; Palmer, C. N. A.; Lyssenko, V.; Tuomi, T.;

Isomaa, B.; Hunter, D. J.; Qi, L.; Shuldiner, A. R.; Roden, M.; Barroso, I.; Wilsaard, T.; Beilby, J.; Hovingh, K.; Price, J. F.; Wilson, J. F.; Rauramaa, R.; Lakka, T. A.; Lind, L.; Dedoussis, G.; Njolstad, I.; Pedersen, N. L.; Khaw, K. T.; Wareham, N. J.; Keinanen-Kiukkaanniemi, S. M.; Saaristo, T. E.; Korpi-Hyovalti, E.; Saltevo, J.; Laakso, M.; Kuusisto, J.; Metspalu, A.; Collins, F. S.; Mohlke, K. L.; Bergman, R. N.; Tuomilehto, J.; Boehm, B. O.; Gieger, C.; Hveem, K.; Cauchi, S.; Froguel, P.; Baldassarre, D.; Tremoli, E.; Humphries, S. E.; Saleheen, D.; Danesh, J.; Ingelsson, E.; Ripatti, S.; Salomaa, V.; Erbel, R.; Jockel, K. H.; Moebus, S.; Peters, A.; Illig, T.; de Faire, U.; Hamsten, A.; Morris, A. D.; Donnelly, P. J.; Frayling, T. M.; Hattersley, A. T.; Boerwinkle, E.; Melander, O.; Kathiresan, S.; Nilsson, P. M.; Deloukas, P.; Thorsteinsdottir, U.; Groop, L. C.; Stefansson, K.; Hu, F.; Pankow, J. S.; Dupuis, J.; Meigs, J. B.; Altshuler, D.; Boehnke, M.; McCarthy, M. I.; Control, W. T. C.; Insulin-Related, M.-A. G.; ANthropometric, G. I.; Epidemiology, A. G.; SAT2D, S. A. T. D.; Replication, D. G., Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **2012**, *44* (9), 981-+.

135. Lango Allen, H.; Estrada, K.; Lettre, G.; Berndt, S. I.; Weedon, M. N.; Rivadeneira, F.; Willer, C. J.; Jackson, A. U.; Vedantam, S.; Raychaudhuri, S.; Ferreira, T.; Wood, A. R.; Weyant, R. J.; Segre, A. V.; Speliotes, E. K.; Wheeler, E.; Soranzo, N.; Park, J. H.; Yang, J.; Gudbjartsson, D.; Heard-Costa, N. L.; Randall, J. C.; Qi, L.; Vernon Smith, A.; Magi, R.; Pastinen, T.; Liang, L.; Heid, I. M.; Luan, J.; Thorleifsson, G.; Winkler, T. W.; Goddard, M. E.; Sin Lo, K.; Palmer, C.; Workalemahu, T.; Aulchenko, Y. S.; Johansson, A.; Zillikens, M. C.; Feitosa, M. F.; Esko, T.; Johnson, T.; Ketkar, S.; Kraft, P.; Mangino, M.; Prokopenko, I.; Absher, D.; Albrecht, E.; Ernst, F.; Glazer, N. L.; Hayward, C.; Hottenga, J. J.; Jacobs, K. B.; Knowles, J. W.; Kutalik, Z.; Monda, K. L.; Polasek, O.; Preuss, M.; Rayner, N. W.; Robertson, N. R.; Steinthorsdottir, V.; Tyrer, J. P.; Voight, B. F.; Wiklund, F.; Xu, J.; Zhao, J. H.; Nyholt, D. R.; Pellikka, N.; Perola, M.; Perry, J. R.; Surakka, I.; Tammesoo, M. L.; Altmaier, E. L.; Amin, N.; Aspelund, T.; Bhangale, T.; Boucher, G.; Chasman, D. I.; Chen, C.; Coin, L.; Cooper, M. N.; Dixon, A. L.; Gibson, Q.; Grundberg, E.; Hao, K.; Juhani Junttila, M.; Kaplan, L. M.; Kettunen, J.; Konig, I. R.; Kwan, T.; Lawrence, R. W.; Levinson, D. F.; Lorentzon, M.; McKnight, B.; Morris, A. P.; Muller, M.; Suh Ngwa, J.; Purcell, S.; Rafelt, S.; Salem, R. M.; Salvi, E.; Sanna, S.; Shi, J.; Sovio, U.; Thompson, J. R.; Turchin, M. C.; Vandenput, L.; Verlaan, D. J.; Vitart, V.; White, C. C.; Ziegler, A.; Almgren, P.; Balmforth, A. J.; Campbell, H.; Citterio, L.; De Grandi, A.; Dominiczak, A.; Duan, J.; Elliott, P.; Elosua, R.; Eriksson, J. G.; Freimer, N. B.; Geus, E. J.; Glorioso, N.; Haiqing, S.; Hartikainen, A. L.; Havulinna, A. S.; Hicks, A. A.; Hui, J.; Igl, W.; Illig, T.; Jula, A.; Kajantie, E.; Kilpelainen, T. O.; Koiranen, M.; Kolcic, I.; Koskinen, S.; Kovacs, P.; Laitinen, J.; Liu, J.; Lokki, M. L.; Marusic, A.; Maschio, A.; Meitinger, T.; Mulas, A.; Pare, G.; Parker, A. N.; Peden, J. F.; Petersmann, A.; Pichler, I.; Pietilainen, K. H.; Pouta, A.; Ridderstrale, M.; Rotter, J. I.; Sambrook, J. G.; Sanders, A. R.; Schmidt, C. O.; Sinisalo, J.; Smit, J. H.; Stringham, H. M.; Bragi Walters, G.; Widen, E.; Wild, S. H.; Willemsen, G.; Zagato, L.; Zgaga, L.; Zitting, P.; Alavere, H.; Farrall, M.; McArdle, W. L.; Nelis, M.; Peters, M. J.; Ripatti, S.; van Meurs, J. B.; Aben, K. K.; Ardlie, K. G.; Beckmann, J. S.; Beilby, J. P.; Bergman, R. N.; Bergmann, S.; Collins, F. S.; Cusi, D.; den Heijer, M.; Eiriksdottir, G.; Gejman, P. V.; Hall, A. S.; Hamsten, A.; Huikuri, H. V.; Iribarren, C.; Kahonen, M.; Kaprio, J.; Kathiresan, S.; Kiemeny, L.; Kocher, T.; Launer, L. J.; Lehtimäki, T.; Melander, O.; Mosley, T. H., Jr.; Musk, A. W.; Nieminen, M. S.; O'Donnell, C. J.; Ohlsson, C.; Oostra, B.; Palmer, L. J.; Raitakari, O.; Ridker, P. M.; Rioux, J. D.; Rissanen, A.; Rivolta, C.; Schunkert, H.; Shuldiner, A. R.; Siscovick, D. S.; Stumvoll, M.; Tonjes, A.; Tuomilehto, J.; van Ommen, G. J.; Viikari, J.; Heath, A. C.; Martin, N. G.; Montgomery, G. W.; Province, M. A.; Kayser, M.; Arnold, A. M.; Atwood, L. D.; Boerwinkle, E.; Chanock, S. J.; Deloukas, P.; Gieger, C.; Gronberg, H.; Hall, P.; Hattersley, A. T.; Hengstenberg, C.; Hoffman, W.; Lathrop, G. M.; Salomaa, V.; Schreiber, S.; Uda, M.; Waterworth, D.; Wright, A. F.; Assimes, T. L.; Barroso, I.; Hofman, A.; Mohlke, K. L.; Boomsma, D. I.; Caulfield, M. J.; Cupples, L. A.; Erdmann, J.; Fox, C. S.; Gudnason, V.; Gyllenstein, U.; Harris, T. B.; Hayes, R. B.; Jarvelin, M. R.; Mooser, V.; Munroe, P. B.; Ouwehand, W. H.; Penninx, B. W.; Pramstaller, P. P.; Quertermous, T.; Rudan, I.; Samani, N. J.; Spector, T. D.; Volzke, H.; Watkins, H.; Wilson, J. F.; Groop, L. C.; Haritunians, T.; Hu, F. B.; Kaplan, R. C.; Metspalu, A.; North, K. E.; Schlessinger, D.; Wareham, N. J.; Hunter, D. J.; O'Connell, J. R.; Strachan, D. P.; Wichmann, H. E.; Borecki, I. B.; van Duijn, C. M.; Schadt, E. E.; Thorsteinsdottir, U.; Peltonen, L.; Uitterlinden, A. G.; Visscher, P. M.; Chatterjee, N.; Loos, R. J.; Boehnke, M.; McCarthy, M. I.; Ingelsson, E.; Lindgren, C. M.; Abecasis, G. R.; Stefansson, K.; Frayling, T. M.; Hirschhorn, J. N., Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **2010**, *467* (7317), 832-8.

136. Wood, A. R.; Esko, T.; Yang, J.; Vedantam, S.; Pers, T. H.; Gustafsson, S.; Chu, A. Y.; Estrada, K.; Luan, J.; Kutalik, Z.; Amin, N.; Buchkovich, M. L.; Croteau-Chonka, D. C.; Day, F. R.; Duan, Y.; Fall, T.; Fehrmann, R.; Ferreira, T.; Jackson, A. U.; Karjalainen, J.; Lo, K. S.; Locke, A. E.; Magi, R.; Mihailov, E.; Porcu, E.; Randall, J. C.; Scherag, A.; Vinkhuyzen, A. A.; Westra, H. J.; Winkler, T. W.; Workalemahu, T.; Zhao, J. H.; Absher, D.; Albrecht, E.; Anderson, D.; Baron, J.; Beekman, M.; Demirkan, A.; Ehret, G. B.; Feenstra, B.; Feitosa, M. F.; Fischer, K.; Fraser, R. M.; Goel, A.; Gong, J.; Justice, A. E.; Kanoni, S.; Kleber, M. E.; Kristiansson, K.; Lim, U.; Lotay, V.; Lui, J. C.; Mangino, M.; Mateo Leach, I.; Medina-Gomez, C.; Nalls, M. A.; Nyholt, D. R.;

- Palmer, C. D.; Pasko, D.; Pechlivanis, S.; Prokopenko, I.; Ried, J. S.; Ripke, S.; Shungin, D.; Stancakova, A.; Strawbridge, R. J.; Sung, Y. J.; Tanaka, T.; Teumer, A.; Trompet, S.; van der Laan, S. W.; van Setten, J.; Van Vliet-Ostaptchouk, J. V.; Wang, Z.; Yengo, L.; Zhang, W.; Afzal, U.; Arnlov, J.; Arscott, G. M.; Bandinelli, S.; Barrett, A.; Bellis, C.; Bennett, A. J.; Berne, C.; Bluher, M.; Bolton, J. L.; Bottcher, Y.; Boyd, H. A.; Bruinenberg, M.; Buckley, B. M.; Buyske, S.; Caspersen, I. H.; Chines, P. S.; Clarke, R.; Claudi-Boehm, S.; Cooper, M.; Daw, E. W.; De Jong, P. A.; Deelen, J.; Delgado, G.; Denny, J. C.; Dhonukshe-Rutten, R.; Dimitriou, M.; Doney, A. S.; Dorr, M.; Eklund, N.; Eury, E.; Folkersen, L.; Garcia, M. E.; Geller, F.; Giedraitis, V.; Go, A. S.; Grallert, H.; Grammer, T. B.; Grassler, J.; Gronberg, H.; de Groot, L. C.; Groves, C. J.; Haessler, J.; Hall, P.; Haller, T.; Hallmans, G.; Hannemann, A.; Hartman, C. A.; Hassinen, M.; Hayward, C.; Heard-Costa, N. L.; Helmer, Q.; Hemani, G.; Henders, A. K.; Hillege, H. L.; Hlatky, M. A.; Hoffmann, W.; Hoffmann, P.; Holmen, O.; Houwing-Duistermaat, J. J.; Illig, T.; Isaacs, A.; James, A. L.; Jeff, J.; Johansen, B.; Johansson, A.; Jolley, J.; Juliusdottir, T.; Junttila, J.; Kho, A. N.; Kinnunen, L.; Klopp, N.; Kocher, T.; Kratzer, W.; Lichtner, P.; Lind, L.; Lindstrom, J.; Lobbens, S.; Lorentzon, M.; Lu, Y.; Lyssenko, V.; Magnusson, P. K.; Mahajan, A.; Maillard, M.; McArdle, W. L.; McKenzie, C. A.; McLachlan, S.; McLaren, P. J.; Menni, C.; Merger, S.; Milani, L.; Moayyeri, A.; Monda, K. L.; Morken, M. A.; Muller, G.; Muller-Nurasyid, M.; Musk, A. W.; Narisu, N.; Nauck, M.; Nolte, I. M.; Nothen, M. M.; Oozageer, L.; Pilz, S.; Rayner, N. W.; Renstrom, F.; Robertson, N. R.; Rose, L. M.; Roussel, R.; Sanna, S.; Scharnagl, H.; Scholtens, S.; Schumacher, F. R.; Schunkert, H.; Scott, R. A.; Sehmi, J.; Seufferlein, T.; Shi, J.; Silventoinen, K.; Smit, J. H.; Smith, A. V.; Smolonska, J.; Stanton, A. V.; Stirrups, K.; Stott, D. J.; Stringham, H. M.; Sundstrom, J.; Swertz, M. A.; Syvanen, A. C.; Tayo, B. O.; Thorleifsson, G.; Tyrer, J. P.; van Dijk, S.; van Schoor, N. M.; van der Velde, N.; van Heemst, D.; van Oort, F. V.; Vermeulen, S. H.; Verweij, N.; Vonk, J. M.; Waite, L. L.; Waldenberger, M.; Wennauer, R.; Wilkens, L. R.; Willenborg, C.; Wilsaard, T.; Wojczynski, M. K.; Wong, A.; Wright, A. F.; Zhang, Q.; Arveiler, D.; Bakker, S. J.; Beilby, J.; Bergman, R. N.; Bergmann, S.; Biffar, R.; Blangero, J.; Boomsma, D. I.; Bornstein, S. R.; Bovet, P.; Brambilla, P.; Brown, M. J.; Campbell, H.; Caulfield, M. J.; Chakravarti, A.; Collins, R.; Collins, F. S.; Crawford, D. C.; Cupples, L. A.; Danesh, J.; de Faire, U.; den Ruijter, H. M.; Erbel, R.; Erdmann, J.; Eriksson, J. G.; Farrall, M.; Ferrannini, E.; Ferrieres, J.; Ford, I.; Forouhi, N. G.; Forrester, T.; Gansevoort, R. T.; Gejman, P. V.; Gieger, C.; Golay, A.; Gottesman, O.; Gudnason, V.; Gyllenstein, U.; Haas, D. W.; Hall, A. S.; Harris, T. B.; Hattersley, A. T.; Heath, A. C.; Hengstenberg, C.; Hicks, A. A.; Hindorf, L. A.; Hingorani, A. D.; Hofman, A.; Hovingh, G. K.; Humphries, S. E.; Hunt, S. C.; Hypponen, E.; Jacobs, K. B.; Jarvelin, M. R.; Jousilahti, P.; Jula, A. M.; Kaprio, J.; Kastelein, J. J.; Kayser, M.; Kee, F.; Keinanen-Kiukkaanniemi, S. M.; Kiemeny, L. A.; Kooner, J. S.; Kooperberg, C.; Koskinen, S.; Kovacs, P.; Kraja, A. T.; Kumari, M.; Kuusisto, J.; Lakka, T. A.; Langenberg, C.; Le Marchand, L.; Lehtimäki, T.; Lupoli, S.; Madden, P. A.; Mannisto, S.; Manunta, P.; Marette, A.; Matise, T. C.; McKnight, B.; Meitinger, T.; Moll, F. L.; Montgomery, G. W.; Morris, A. D.; Morris, A. P.; Murray, J. C.; Nelis, M.; Ohlsson, C.; Oldehinkel, A. J.; Ong, K. K.; Ouweland, W. H.; Pasterkamp, G.; Peters, A.; Pramstaller, P. P.; Price, J. F.; Qi, L.; Raitakari, O. T.; Rankinen, T.; Rao, D. C.; Rice, T. K.; Ritchie, M.; Rudan, I.; Salomaa, V.; Samani, N. J.; Saramies, J.; Sarzynski, M. A.; Schwarz, P. E.; Sebert, S.; Sever, P.; Shuldiner, A. R.; Sinisalo, J.; Steinthorsdottir, V.; Stolk, R. P.; Tardif, J. C.; Tonjes, A.; Tremblay, A.; Tremoli, E.; Virtamo, J.; Vohl, M. C.; Electronic Medical, R.; Genomics, C.; Consortium, M. I.; Consortium, P.; LifeLines Cohort, S.; Amouyel, P.; Asselbergs, F. W.; Assimes, T. L.; Bochud, M.; Boehm, B. O.; Boerwinkle, E.; Bottinger, E. P.; Bouchard, C.; Cauchi, S.; Chambers, J. C.; Chanock, S. J.; Cooper, R. S.; de Bakker, P. I.; Dedoussis, G.; Ferrucci, L.; Franks, P. W.; Froguel, P.; Groop, L. C.; Haiman, C. A.; Hamsten, A.; Hayes, M. G.; Hui, J.; Hunter, D. J.; Hveem, K.; Jukema, J. W.; Kaplan, R. C.; Kivimäki, M.; Kuh, D.; Laakso, M.; Liu, Y.; Martin, N. G.; Marz, W.; Melbye, M.; Moebus, S.; Munroe, P. B.; Njolstad, I.; Oostra, B. A.; Palmer, C. N.; Pedersen, N. L.; Perola, M.; Perusse, L.; Peters, U.; Powell, J. E.; Power, C.; Quertermous, T.; Rauramaa, R.; Reinmaa, E.; Ridker, P. M.; Rivadeneira, F.; Rotter, J. I.; Saaristo, T. E.; Saleheen, D.; Schlessinger, D.; Slagboom, P. E.; Snieder, H.; Spector, T. D.; Strauch, K.; Stumvoll, M.; Tuomilehto, J.; Uusitupa, M.; van der Harst, P.; Volzke, H.; Walker, M.; Wareham, N. J.; Watkins, H.; Wichmann, H. E.; Wilson, J. F.; Zanen, P.; Deloukas, P.; Heid, I. M.; Lindgren, C. M.; Mohlke, K. L.; Speliotes, E. K.; Thorsteinsdottir, U.; Barroso, I.; Fox, C. S.; North, K. E.; Strachan, D. P.; Beckmann, J. S.; Berndt, S. I.; Boehnke, M.; Borecki, I. B.; McCarthy, M. I.; Metspalu, A.; Stefansson, K.; Uitterlinden, A. G.; van Duijn, C. M.; Franke, L.; Willer, C. J.; Price, A. L.; Lettre, G.; Loos, R. J.; Weedon, M. N.; Ingelsson, E.; O'Connell, J. R.; Abecasis, G. R.; Chasman, D. I.; Goddard, M. E.; Visscher, P. M.; Hirschhorn, J. N.; Frayling, T. M., Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **2014**, *46* (11), 1173-86.
137. Pers, T. H.; Karjalainen, J. M.; Chan, Y.; Westra, H. J.; Wood, A. R.; Yang, J.; Lui, J. C.; Vedantam, S.; Gustafsson, S.; Esko, T.; Frayling, T.; Speliotes, E. K.; Genetic Investigation of, A. T. C.; Boehnke, M.; Raychaudhuri, S.; Fehrmann, R. S.; Hirschhorn, J. N.; Franke, L., Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **2015**, *6*, 5890.
138. Klingseisen, A.; Jackson, A. P., Mechanisms and pathways of growth failure in primordial dwarfism. *Genes Dev* **2011**, *25* (19), 2011-24.

139. Bartholdi, D.; Krajewska-Walasek, M.; Ounap, K.; Gaspar, H.; Chrzanowska, K. H.; Ilyana, H.; Kayserili, H.; Lurie, I. W.; Schinzel, A.; Baumer, A., Epigenetic mutations of the imprinted IGF2-H19 domain in Silver-Russell syndrome (SRS): results from a large cohort of patients with SRS and SRS-like phenotypes. *J Med Genet* **2009**, *46* (3), 192-7.
140. Marouli, E.; Graff, M.; Medina-Gomez, C.; Lo, K. S.; Wood, A. R.; Kjaer, T. R.; Fine, R. S.; Lu, Y.; Schurmann, C.; Highland, H. M.; Rueger, S.; Thorleifsson, G.; Justice, A. E.; Lamparter, D.; Stirrups, K. E.; Turcot, V.; Young, K. L.; Winkler, T. W.; Esko, T.; Karaderi, T.; Locke, A. E.; Masca, N. G.; Ng, M. C.; Mudgal, P.; Rivas, M. A.; Vedantam, S.; Mahajan, A.; Guo, X.; Abecasis, G.; Aben, K. K.; Adair, L. S.; Alam, D. S.; Albrecht, E.; Allin, K. H.; Allison, M.; Amouyel, P.; Appel, E. V.; Arveiler, D.; Asselbergs, F. W.; Auer, P. L.; Balkau, B.; Banas, B.; Bang, L. E.; Benn, M.; Bergmann, S.; Bielak, L. F.; Bluher, M.; Boeing, H.; Boerwinkle, E.; Boger, C. A.; Bonnyycastle, L. L.; Bork-Jensen, J.; Bots, M. L.; Bottinger, E. P.; Bowden, D. W.; Brandslund, I.; Breen, G.; Brilliant, M. H.; Broer, L.; Burt, A. A.; Butterworth, A. S.; Carey, D. J.; Caulfield, M. J.; Chambers, J. C.; Chasman, D. I.; Chen, Y. I.; Chowdhury, R.; Christensen, C.; Chu, A. Y.; Cocca, M.; Collins, F. S.; Cook, J. P.; Corley, J.; Galbany, J. C.; Cox, A. J.; Cuellar-Partida, G.; Danesh, J.; Davies, G.; de Bakker, P. I.; de Borst, G. J.; de Denu, S.; de Groot, M. C.; de Mutsert, R.; Deary, I. J.; Dedoussis, G.; Demerath, E. W.; den Hollander, A. I.; Dennis, J. G.; Di Angelantonio, E.; Drenos, F.; Du, M.; Dunning, A. M.; Easton, D. F.; Ebeling, T.; Edwards, T. L.; Ellinor, P. T.; Elliott, P.; Evangelou, E.; Farmaki, A. E.; Faul, J. D.; Feitosa, M. F.; Feng, S.; Ferrannini, E.; Ferrario, M. M.; Ferrieres, J.; Florez, J. C.; Ford, I.; Fornage, M.; Franks, P. W.; Frikke-Schmidt, R.; Galesloot, T. E.; Gan, W.; Gandini, P.; Gasparini, P.; Giedraitis, V.; Giri, A.; Girotto, G.; Gordon, S. D.; Gordon-Larsen, P.; Gorski, M.; Grarup, N.; Grove, M. L.; Gudnason, V.; Gustafsson, S.; Hansen, T.; Harris, K. M.; Harris, T. B.; Hattersley, A. T.; Hayward, C.; He, L.; Heid, I. M.; Heikkila, K.; Helgeland, O.; Hernesniemi, J.; Hewitt, A. W.; Hocking, L. J.; Hollensted, M.; Holmen, O. L.; Hovingh, G. K.; Howson, J. M.; Hoyng, C. B.; Huang, P. L.; Hveem, K.; Ikram, M. A.; Ingelsson, E.; Jackson, A. U.; Jansson, J. H.; Jarvik, G. P.; Jensen, G. B.; Jhun, M. A.; Jia, Y.; Jiang, X.; Johansson, S.; Jorgensen, M. E.; Jorgensen, T.; Jousilahti, P.; Jukema, J. W.; Kahali, B.; Kahn, R. S.; Kahonen, M.; Kamstrup, P. R.; Kanoni, S.; Kaprio, J.; Karaleftheri, M.; Kardina, S. L.; Karpe, F.; Kee, F.; Keeman, R.; Kiemeny, L. A.; Kitajima, H.; Kluivers, K. B.; Kocher, T.; Komulainen, P.; Kontto, J.; Kooner, J. S.; Kooperberg, C.; Kovacs, P.; Kriebel, J.; Kuivaniemi, H.; Kury, S.; Kuusisto, J.; La Bianca, M.; Laakso, M.; Lakka, T. A.; Lange, E. M.; Lange, L. A.; Langefeld, C. D.; Langenberg, C.; Larson, E. B.; Lee, I. T.; Lehtimäki, T.; Lewis, C. E.; Li, H.; Li, J.; Li-Gao, R.; Lin, H.; Lin, L. A.; Lin, X.; Lind, L.; Lindstrom, J.; Linneberg, A.; Liu, Y.; Liu, Y.; Lophatananon, A.; Luan, J.; Lubitz, S. A.; Lyytikäinen, L. P.; Mackey, D. A.; Madden, P. A.; Manning, A. K.; Mannisto, S.; Marenne, G.; Marten, J.; Martin, N. G.; Mazul, A. L.; Meidtner, K.; Metspalu, A.; Mitchell, P.; Mohlke, K. L.; Mook-Kanamori, D. O.; Morgan, A.; Morris, A. D.; Morris, A. P.; Muller-Nurasyid, M.; Munroe, P. B.; Nalls, M. A.; Nauck, M.; Nelson, C. P.; Neville, M.; Nielsen, S. F.; Nikus, K.; Njolstad, P. R.; Nordestgaard, B. G.; Ntalla, I.; O'Connel, J. R.; Oksa, H.; Loohuis, L. M.; Ophoff, R. A.; Owen, K. R.; Packard, C. J.; Padmanabhan, S.; Palmer, C. N.; Pasterkamp, G.; Patel, A. P.; Pattie, A.; Pedersen, O.; Peissig, P. L.; Peloso, G. M.; Pennell, C. E.; Perola, M.; Perry, J. A.; Perry, J. R.; Person, T. N.; Pirie, A.; Polasek, O.; Posthuma, D.; Raitakari, O. T.; Rasheed, A.; Rauramaa, R.; Reilly, D. F.; Reiner, A. P.; Renstrom, F.; Ridker, P. M.; Rioux, J. D.; Robertson, N.; Robino, A.; Rolandsson, O.; Rudan, I.; Ruth, K. S.; Saleheen, D.; Salomaa, V.; Samani, N. J.; Sadow, K.; Sapkota, Y.; Sattar, N.; Schmidt, M. K.; Schreiner, P. J.; Schulze, M. B.; Scott, R. A.; Segura-Lepe, M. P.; Shah, S.; Sim, X.; Sivapalaratnam, S.; Small, K. S.; Smith, A. V.; Smith, J. A.; Southam, L.; Spector, T. D.; Speliotes, E. K.; Starr, J. M.; Steinthorsdottir, V.; Stringham, H. M.; Stumvoll, M.; Surendran, P.; t Hart, L. M.; Tansey, K. E.; Tardif, J. C.; Taylor, K. D.; Teumer, A.; Thompson, D. J.; Thorsteinsdottir, U.; Thuesen, B. H.; Tonjes, A.; Tromp, G.; Trompet, S.; Tsaftakis, E.; Tuomilehto, J.; Tybjaerg-Hansen, A.; Tyrer, J. P.; Uher, R.; Uitterlinden, A. G.; Ulivi, S.; van der Laan, S. W.; Van Der Leij, A. R.; van Duijn, C. M.; van Schoor, N. M.; van Setten, J.; Varbo, A.; Varga, T. V.; Varma, R.; Edwards, D. R.; Vermeulen, S. H.; Vestergaard, H.; Vitart, V.; Vogt, T. F.; Vozzi, D.; Walker, M.; Wang, F.; Wang, C. A.; Wang, S.; Wang, Y.; Wareham, N. J.; Warren, H. R.; Wessel, J.; Willems, S. M.; Wilson, J. G.; Witte, D. R.; Woods, M. O.; Wu, Y.; Yaghootkar, H.; Yao, J.; Yao, P.; Yerges-Armstrong, L. M.; Young, R.; Zeggini, E.; Zhan, X.; Zhang, W.; Zhao, J. H.; Zhao, W.; Zhao, W.; Zheng, H.; Zhou, W.; Consortium, E. P.-I.; Consortium, C. H. D. E.; Exome, B. P. C.; Consortium, T. D.-G.; Go, T. D. G. C.; Global Lipids Genetics, C.; ReproGen, C.; Investigators, M.; Rotter, J. I.; Boehnke, M.; Kathiresan, S.; McCarthy, M. I.; Willer, C. J.; Stefansson, K.; Borecki, I. B.; Liu, D. J.; North, K. E.; Heard-Costa, N. L.; Pers, T. H.; Lindgren, C. M.; Oxvig, C.; Kutalik, Z.; Rivadeneira, F.; Loos, R. J.; Frayling, T. M.; Hirschhorn, J. N.; Deloukas, P.; Lettre, G., Rare and low-frequency coding variants alter human adult height. *Nature* **2017**, *542* (7640), 186-190.
141. Schwartz, N. B.; Domowicz, M., Chondrodysplasias due to proteoglycan defects. *Glycobiology* **2002**, *12* (4), 57r-68r.
142. Kim, H.; Kim, I. Y.; Lee, S. Y.; Jeong, D., Bimodal actions of reactive oxygen species in the differentiation and bone-resorbing functions of osteoclasts. *FEBS Lett* **2006**, *580* (24), 5661-5.

143. Smith, L. B.; Belanger, J. M.; Oberbauer, A. M., Fibroblast growth factor receptor 3 effects on proliferation and telomerase activity in sheep growth plate chondrocytes. *J Anim Sci Biotechnol* **2012**, *3* (1), 39.
144. Cho, Y. S.; Go, M. J.; Kim, Y. J.; Heo, J. Y.; Oh, J. H.; Ban, H. J.; Yoon, D.; Lee, M. H.; Kim, D. J.; Park, M.; Cha, S. H.; Kim, J. W.; Han, B. G.; Min, H.; Ahn, Y.; Park, M. S.; Han, H. R.; Jang, H. Y.; Cho, E. Y.; Lee, J. E.; Cho, N. H.; Shin, C.; Park, T.; Park, J. W.; Lee, J. K.; Cardon, L.; Clarke, G.; McCarthy, M. I.; Lee, J. Y.; Lee, J. K.; Oh, B.; Kim, H. L., A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* **2009**, *41* (5), 527-34.
145. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C., STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* **2015**, *43* (Database issue), D447-52.
146. Safran, M.; Dalah, I.; Alexander, J.; Rosen, N.; Iny Stein, T.; Shmoish, M.; Nativ, N.; Bahir, I.; Doniger, T.; Krug, H.; Sirota-Madi, A.; Olender, T.; Golan, Y.; Stelzer, G.; Harel, A.; Lancet, D., GeneCards Version 3: the human gene integrator. *Database (Oxford)* **2010**, *2010*, baq020.
147. Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K., dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **2001**, *29* (1), 308-11.
148. Tessneer, K. L.; Jackson, R. M.; Griesel, B. A.; Olson, A. L., Rab5 Activity Regulates GLUT4 Sorting Into Insulin-Responsive and Non-Insulin-Responsive Endosomal Compartments: A Potential Mechanism for Development of Insulin Resistance. *Endocrinology* **2014**, *155* (9), 3315-3328.
149. Huang, J.; Imamura, T.; Olefsky, J. M., Insulin can regulate GLUT4 internalization by signaling to Rab5 and the motor protein dynein. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98* (23), 13084-13089.
150. Moller, D. E., Potential role of TNF-alpha in the pathogenesis of insulin resistance and type 2 diabetes. *Trends in Endocrinology and Metabolism* **2000**, *11* (6), 212-217.
151. Huang, S. H.; Czech, M. P., The GLUT4 glucose transporter. *Cell Metabolism* **2007**, *5* (4), 237-252.
152. Gaster, M.; Staehr, P.; Beck-Nielsen, H.; Schroder, H. D.; Handberg, A., GLUT4 is reduced in slow muscle fibers of type 2 diabetic patients - Is insulin resistance in type 2 diabetes a slow, type 1 fiber disease? *Diabetes* **2001**, *50* (6), 1324-1329.
153. Krakow, D.; Rimoim, D. L., The skeletal dysplasias. *Genet Med* **2010**, *12* (6), 327-41.
154. Simeone, P.; Alberti, S., Epigenetic heredity of human height. *Physiol Rep* **2014**, *2* (6).
155. Pappas, J. G., The Clinical Course of an Overgrowth Syndrome, From Diagnosis in Infancy Through Adulthood: The Case of Beckwith-Wiedemann Syndrome. *Curr Prob Pediatr Ad* **2015**, *45* (4), 112-117.
156. Faravelli, F., NSD1 mutations in Sotos syndrome. *Am J Med Genet C Semin Med Genet* **2005**, *137C* (1), 24-31.
157. Laron, Z., Insulin-like growth factor 1 (IGF-1): a growth hormone. *Mol Pathol* **2001**, *54* (5), 311-6.
158. Malhotra, D.; Yang, Y., Wnts' fashion statement: from body stature to dysplasia. *Bonekey Rep* **2014**, *3*, 541.
159. Le Goff, C.; Cormier-Daire, V., Chondrodysplasias and TGFbeta signaling. *Bonekey Rep* **2015**, *4*, 642.
160. Olney, R. C.; Wang, J.; Sylvester, J. E.; Mougey, E. B., Growth factor regulation of human growth plate chondrocyte proliferation in vitro. *Biochem Biophys Res Commun* **2004**, *317* (4), 1171-82.
161. Myllyharju, J., Extracellular matrix and developing growth plate. *Curr Osteoporos Rep* **2014**, *12* (4), 439-45.
162. Dobрева, G.; Chahrouh, M.; Dautzenberg, M.; Chirivella, L.; Kanzler, B.; Farinas, I.; Karsenty, G.; Grosschedl, R., SATB2 is a multifunctional determinant of craniofacial patterning and osteoblast differentiation. *Cell* **2006**, *125* (5), 971-86.
163. Zemel, B. S.; Katz, S. H., The contribution of adrenal and gonadal androgens to the growth in height of adolescent males. *Am J Phys Anthropol* **1986**, *71* (4), 459-66.
164. Severe, N.; Dieudonne, F. X.; Marie, P. J., E3 ubiquitin ligase-mediated regulation of bone formation and tumorigenesis. *Cell Death Dis* **2013**, *4*, e463.
165. De Luca, F., Role of Nuclear Factor Kappa B (NF-kappaB) in Growth Plate Chondrogenesis. *Pediatr Endocrinol Rev* **2016**, *13* (4), 720-30.
166. Chen, K.; Rajewsky, N., The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **2007**, *8* (2), 93-103.
167. Salmena, L.; Poliseno, L.; Tay, Y.; Kats, L.; Pandolfi, P. P., A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **2011**, *146* (3), 353-8.
168. Bueno, M. J.; Malumbres, M., MicroRNAs and the cell cycle. *Biochim Biophys Acta* **2011**, *1812* (5), 592-601.

169. Shivdasani, R. A., MicroRNAs: regulators of gene expression and cell differentiation. *Blood* **2006**, *108* (12), 3646-53.
170. Neal, C. S.; Michael, M. Z.; Pimlott, L. K.; Yong, T. Y.; Li, J. Y. Z.; Gleadle, J. M., Circulating microRNA expression is reduced in chronic kidney disease. *Nephrol Dial Transpl* **2011**, *26* (11), 3794-3802.
171. Zhang, B. H.; Pan, X. P.; Cobb, G. P.; Anderson, T. A., microRNAs as oncogenes and tumor suppressors. *Dev Biol* **2007**, *302* (1), 1-12.
172. John, B.; Enright, A. J.; Aravin, A.; Tuschl, T.; Sander, C.; Marks, D. S., Human MicroRNA targets. *PLoS Biol* **2004**, *2* (11), e363.
173. Lewis, B. P.; Burge, C. B.; Bartel, D. P., Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **2005**, *120* (1), 15-20.
174. Krek, A.; Grun, D.; Poy, M. N.; Wolf, R.; Rosenberg, L.; Epstein, E. J.; MacMenamin, P.; da Piedade, I.; Gunsalus, K. C.; Stoffel, M.; Rajewsky, N., Combinatorial microRNA target predictions. *Nat Genet* **2005**, *37* (5), 495-500.
175. Kertesz, M.; Iovino, N.; Unnerstall, U.; Gaul, U.; Segal, E., The role of site accessibility in microRNA target recognition. *Nat Genet* **2007**, *39* (10), 1278-84.
176. Kiriakidou, M.; Nelson, P. T.; Kouranov, A.; Fitziev, P.; Bouyioukos, C.; Mourelatos, Z.; Hatzigeorgiou, A., A combined computational-experimental approach predicts human microRNA targets. *Gene Dev* **2004**, *18* (10), 1165-1178.
177. Huang, J. C.; Babak, T.; Corson, T. W.; Chua, G.; Khan, S.; Gallie, B. L.; Hughes, T. R.; Blencowe, B. J.; Frey, B. J.; Morris, Q. D., Using expression profiling data to identify human microRNA targets. *Nature Methods* **2007**, *4* (12), 1045-1049.
178. Lu, Y.; Zhou, Y.; Qu, W.; Deng, M.; Zhang, C., A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics (Oxford, England)* **2011**, *27* (17), 2406-13.
179. Muniategui, A.; Pey, J.; Planes, F. J.; Rubio, A., Joint analysis of miRNA and mRNA expression data. *Briefings in bioinformatics* **2013**, *14* (3), 263-78.
180. Yoon, S.; De Micheli, G., Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics (Oxford, England)* **2005**, *21* Suppl 2, ii93-100.
181. Pio, G.; Ceci, M.; D'Elia, D.; Loglisci, C.; Malerba, D., A novel biclustering algorithm for the discovery of meaningful biological correlations between microRNAs and their target genes. *BMC Bioinformatics* **2013**, *14* Suppl 7, S8.
182. Joung, J. G.; Hwang, K. B.; Nam, J. W.; Kim, S. J.; Zhang, B. T., Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics* **2007**, *23* (9), 1141-1147.
183. Peng, X.; Li, Y.; Walters, K. A.; Rosenzweig, E. R.; Lederer, S. L.; Aicher, L. D.; Prohl, S.; Katze, M. G., Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC genomics* **2009**, *10*, 373.
184. Liu, B.; Li, J.; Cairns, M. J., Identifying miRNAs, targets and functions. *Briefings in bioinformatics* **2014**, *15* (1), 1-19.
185. Liu, B.; Liu, L.; Tsykin, A.; Goodall, G. J.; Green, J. E.; Zhu, M.; Kim, C. H.; Li, J., Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics (Oxford, England)* **2010**, *26* (24), 3105-11.
186. Mitra, K.; Carvunis, A. R.; Ramesh, S. K.; Ideker, T., Integrative approaches for finding modular structure in biological networks. *Nature reviews. Genetics* **2013**, *14* (10), 719-32.
187. Knudsen, S.; Hother, C.; Gronbaek, K.; Jensen, T.; Hansen, A.; Mazin, W.; Dahlgaard, J.; Moller, M. B.; Ralfkiaer, E.; Brown, P. D., Development and Blind Clinical Validation of a MicroRNA Based Predictor of Response to Treatment with R-CHO(E)P in DLBCL. *Plos One* **2015**, *10* (2).
188. Clough, E.; Barrett, T., The Gene Expression Omnibus Database. *Methods Mol Biol* **2016**, *1418*, 93-110.
189. Gennarino, V. A.; D'Angelo, G.; Dharmalingam, G.; Fernandez, S.; Russolillo, G.; Sanges, R.; Mutarelli, M.; Belcastro, V.; Ballabio, A.; Verde, P.; Sardiello, M.; Banfi, S., Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res* **2012**, *22* (6), 1163-72.
190. Bondy, J. A.; Murty, U. S. R., *Graph theory with applications*. Macmillan: London, 1976; p x, 264 p.
191. Prelic, A.; Bleuler, S.; Zimmermann, P.; Wille, A.; Buhlmann, P.; Gruissem, W.; Hennig, L.; Thiele, L.; Zitzler, E., A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **2006**, *22* (9), 1122-9.
192. Bergmann, S.; Ihmels, J.; Barkai, N., Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E* **2003**, *67* (3).
193. Li, G.; Ma, Q.; Tang, H.; Paterson, A. H.; Xu, Y., QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res* **2009**, *37* (15), e101.

194. Hochreiter, S.; Bodenhofer, U.; Heusel, M.; Mayr, A.; Mitterecker, A.; Kasim, A.; Khamiakova, T.; Van Sanden, S.; Lin, D.; Talloen, W.; Bijnsens, L.; Gohlmann, H. W.; Shkedy, Z.; Clevert, D. A., FABIA: factor analysis for bicluster acquisition. *Bioinformatics* **2010**, *26* (12), 1520-7.
195. Rodriguez-Baena, D. S.; Perez-Pulido, A. J.; Aguilar-Ruiz, J. S., A biclustering algorithm for extracting bit-patterns from binary datasets. *Bioinformatics* **2011**, *27* (19), 2738-45.
196. Edgar, R.; Domrachev, M.; Lash, A. E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **2002**, *30* (1), 207-10.
197. Gautier, L.; Cope, L.; Bolstad, B. M.; Irizarry, R. A., affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **2004**, *20* (3), 307-315.
198. Garcia, D. M.; Baek, D.; Shin, C.; Bell, G. W.; Grimson, A.; Bartel, D. P., Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol* **2011**, *18* (10), 1139-46.
199. Kozomara, A.; Griffiths-Jones, S., miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **2014**, *42* (Database issue), D68-73.
200. Betel, D.; Koppal, A.; Agius, P.; Sander, C.; Leslie, C., Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* **2010**, *11* (8), R90.
201. Kertesz, M.; Iovino, N.; Unnerstall, U.; Gaul, U.; Segal, E., The role of site accessibility in microRNA target recognition. *Nat Genet* **2007**, *39* (10), 1278-1284.
202. Maragkakis, M.; Reczko, M.; Simossis, V. A.; Alexiou, P.; Papadopoulos, G. L.; Dalamagas, T.; Giannopoulos, G.; Goumas, G.; Koukis, E.; Kourtis, K.; Vergoulis, T.; Koziris, N.; Sellis, T.; Tsanakas, P.; Hatzigeorgiou, A. G., DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research* **2009**, *37*, W273-W276.
203. Paraskevopoulou, M. D.; Georgakilas, G.; Kostoulas, N.; Vlachos, I. S.; Vergoulis, T.; Reczko, M.; Filippidis, C.; Dalamagas, T.; Hatzigeorgiou, A. G., DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research* **2013**, *41* (W1), W169-W173.
204. Wang, X. W., Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics* **2016**, *32* (9), 1316-1322.
205. Nielsen, C. B.; Shomron, N.; Sandberg, R.; Hornstein, E.; Kitzman, J.; Burge, C. B., Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *Rna* **2007**, *13* (11), 1894-1910.
206. Li, G. J.; Ma, Q.; Tang, H. B.; Paterson, A. H.; Xu, Y., QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research* **2009**, *37* (15).
207. Chou, C. H.; Shrestha, S.; Yang, C. D.; Chang, N. W.; Lin, Y. L.; Liao, K. W.; Huang, W. C.; Sun, T. H.; Tu, S. J.; Lee, W. H.; Chiew, M. Y.; Tai, C. S.; Wei, T. Y.; Tsai, T. R.; Huang, H. T.; Wang, C. Y.; Wu, H. Y.; Ho, S. Y.; Chen, P. R.; Chuang, C. H.; Hsieh, P. J.; Wu, Y. S.; Chen, W. L.; Li, M. J.; Wu, Y. C.; Huang, X. Y.; Ng, F. L.; Buddhakosai, W.; Huang, P. C.; Lan, K. C.; Huang, C. Y.; Weng, S. L.; Cheng, Y. N.; Liang, C.; Hsu, W. L.; Huang, H. D., miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* **2018**, *46* (D1), D296-D302.
208. Wang, P.; Ning, S. W.; Wang, Q. H.; Li, R. H.; Ye, J. R.; Zhao, Z. X. L.; Li, Y.; Huang, T.; Li, X., mirTarPri: Improved Prioritization of MicroRNA Targets through Incorporation of Functional Genomics Data. *Plos One* **2013**, *8* (1).
209. Santosa, F.; Symes, W. W., Linear Inversion of Band-Limited Reflection Seismograms. *Siam J Sci Stat Comp* **1986**, *7* (4), 1307-1330.
210. Tibshirani, R., Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* **1996**, *58* (1), 267-288.
211. Sass, S.; Pitea, A.; Unger, K.; Hess, J.; Mueller, N. S.; Theis, F. J., MicroRNA-Target Network Inference and Local Network Enrichment Analysis Identify Two microRNA Clusters with Distinct Functions in Head and Neck Squamous Cell Carcinoma. *Int J Mol Sci* **2015**, *16* (12), 30204-22.
212. Le, T. D.; Liu, L.; Tsykin, A.; Goodall, G. J.; Liu, B.; Sun, B. Y.; Li, J., Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics* **2013**, *29* (6), 765-71.
213. Koo, J.; Zhang, J. Y.; Chaterji, S., Tiresias: Context-sensitive Approach to Decipher the Presence and Strength of MicroRNA Regulatory Interactions. *Theranostics* **2018**, *8* (1), 277-291.
214. Le, T. D.; Zhang, J.; Liu, L.; Liu, H.; Li, J., miRLAB: An R Based Dry Lab for Exploring miRNA-mRNA Regulatory Relationships. *PLoS One* **2015**, *10* (12), e0145386.
215. Chang, F.; Lee, J. T.; Navolanic, P. M.; Steelman, L. S.; Shelton, J. G.; Blalock, W. L.; Franklin, R. A.; McCubrey, J. A., Involvement of PI3K/Akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy. *Leukemia* **2003**, *17* (3), 590-603.
216. Luo, J.; Manning, B. D.; Cantley, L. C., Targeting the PI3K-Akt pathway in human cancer: rationale and promise. *Cancer Cell* **2003**, *4* (4), 257-62.

217. Chou, J.; Lin, J. H.; Brenot, A.; Kim, J. W.; Provot, S.; Werb, Z., GATA3 suppresses metastasis and modulates the tumour microenvironment by regulating microRNA-29b expression. *Nat Cell Biol* **2013**, *15* (2), 201-13.
218. Buffa, F. M.; Camps, C.; Winchester, L.; Snell, C. E.; Gee, H. E.; Sheldon, H.; Taylor, M.; Harris, A. L.; Ragoussis, J., microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Res* **2011**, *71* (17), 5635-45.
219. GeneCards. www.genecards.org.
220. Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M.; Yefanov, A.; Lee, H.; Zhang, N.; Robertson, C. L.; Serova, N.; Davis, S.; Soboleva, A., NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **2013**, *41* (Database issue), D991-5.
221. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **2015**, *348* (6235), 648-60.
222. Wang, Y. P.; Li, K. B., Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *Bmc Genomics* **2009**, *10*.
223. Ng, E. K.; Tsang, W. P.; Ng, S. S.; Jin, H. C.; Yu, J.; Li, J. J.; Rocken, C.; Ebert, M. P.; Kwok, T. T.; Sung, J. J., MicroRNA-143 targets DNA methyltransferases 3A in colorectal cancer. *Br J Cancer* **2009**, *101* (4), 699-706.
224. Bryan, K.; Terrile, M.; Bray, I. M.; Domingo-Fernandez, R.; Watters, K. M.; Koster, J.; Versteeg, R.; Stallings, R. L., Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis. *Nucleic Acids Res* **2014**, *42* (3), e17.
225. Reiss, D. J.; Plaisier, C. L.; Wu, W. J.; Baliga, N. S., cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism. *Nucleic Acids Res* **2015**, *43* (13), e87.
226. Gonzalez-Dominguez, J.; Exposit, R. R., ParBiBit: Parallel tool for binary biclustering on modern distributed-memory systems. *Plos One* **2018**, *13* (4).
227. Worringer, K. A.; Rand, T. A.; Hayashi, Y.; Sami, S.; Takahashi, K.; Tanabe, K.; Narita, M.; Srivastava, D.; Yamanaka, S., The let-7/LIN-41 pathway regulates reprogramming to human induced pluripotent stem cells by controlling expression of prodifferentiation genes. *Cell Stem Cell* **2014**, *14* (1), 40-52.
228. Rahkonen, N.; Stubb, A.; Malonzo, M.; Edelman, S.; Emani, M. R.; Narva, E.; Lahdesmaki, H.; Ruohola-Baker, H.; Lahesmaa, R.; Lund, R., Mature Let-7 miRNAs fine tune expression of LIN28B in pluripotent human embryonic stem cells. *Stem Cell Res* **2016**, *17* (3), 498-503.
229. Zipeto, M. A.; Court, A. C.; Sadarangani, A.; Delos Santos, N. P.; Balaian, L.; Chun, H. J.; Pineda, G.; Morris, S. R.; Mason, C. N.; Geron, I.; Barrett, C.; Goff, D. J.; Wall, R.; Pellicchia, M.; Minden, M.; Frazer, K. A.; Marra, M. A.; Crews, L. A.; Jiang, Q.; Jamieson, C. H. M., ADAR1 Activation Drives Leukemia Stem Cell Self-Renewal by Impairing Let-7 Biogenesis. *Cell Stem Cell* **2016**, *19* (2), 177-191.
230. Liao, T. T.; Yang, M. H., Downregulation of Let-7i to promote stem-like properties of head and neck cancer cells through activating ARID3B-Oct4 axis. *J Clin Oncol* **2013**, *31* (15).
231. Johnson, C. D.; Esquela-Kerscher, A.; Stefani, G.; Byrom, M.; Kelnar, K.; Ovcharenko, D.; Wilson, M.; Wang, X.; Shelton, J.; Shingara, J.; Chin, L.; Brown, D.; Slack, F. J., The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer Res* **2007**, *67* (16), 7713-22.
232. Hu, X.; Guo, J.; Zheng, L.; Li, C.; Zheng, T. M.; Tanyi, J. L.; Liang, S.; Benedetto, C.; Mitidieri, M.; Katsaros, D.; Zhao, X.; Zhang, Y.; Huang, Q.; Zhang, L., The heterochronic microRNA let-7 inhibits cell motility by regulating the genes in the actin cytoskeleton pathway in breast cancer. *Mol Cancer Res* **2013**, *11* (3), 240-50.
233. Chafin, C. B.; Regna, N. L.; Caudell, D. L.; Reilly, C. M., MicroRNA-let-7a promotes E2F-mediated cell proliferation and NFkappaB activation in vitro. *Cell Mol Immunol* **2014**, *11* (1), 79-83.
234. Liu, K.; Zhang, C.; Li, T.; Ding, Y.; Tu, T.; Zhou, F.; Qi, W.; Chen, H.; Sun, X., Let-7a inhibits growth and migration of breast cancer cells by targeting HMGA1. *Int J Oncol* **2015**, *46* (6), 2526-34.
235. Boyerinas, B.; Park, S. M.; Shomron, N.; Hedegaard, M. M.; Vinther, J.; Andersen, J. S.; Feig, C.; Xu, J.; Burge, C. B.; Peter, M. E., Identification of let-7-regulated oncofetal genes. *Cancer Res* **2008**, *68* (8), 2587-91.
236. Rybak, A.; Fuchs, H.; Smirnova, L.; Brandt, C.; Pohl, E. E.; Nitsch, R.; Wulczyn, F. G., A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. *Nat Cell Biol* **2008**, *10* (8), 987-93.
237. Elkhadragy, L.; Chen, M.; Miller, K.; Yang, M. H.; Long, W., A regulatory BMI1/let-7i/ERK3 pathway controls the motility of head and neck cancer cells. *Mol Oncol* **2017**, *11* (2), 194-207.
238. Powers, J. T.; Tsanov, K. M.; Pearson, D. S.; Roels, F.; Spina, C. S.; Ebright, R.; Seligson, M.; de Soysa, Y.; Cahan, P.; Theissen, J.; Tu, H. C.; Han, A.; Kurek, K. C.; LaPier, G. S.; Osborne, J. K.; Ross, S. J.; Cesana, M.; Collins, J. J.; Berthold, F.; Daley, G. Q., Multiple mechanisms disrupt the let-7 microRNA family in neuroblastoma. *Nature* **2016**, *535* (7611), 246-51.

239. Chen, K. C.; Hsieh, I. C.; Hsi, E.; Wang, Y. S.; Dai, C. Y.; Chou, W. W.; Juo, S. H. H., Negative feedback regulation between microRNA let-7g and the oxLDL receptor LOX-1. *J Cell Sci* **2011**, *124* (23), 4115-4124.
240. Russ, A. C.; Sander, S.; Luck, S. C.; Lang, K. M.; Bauer, M.; Rucker, F. G.; Kestler, H. A.; Schlenk, R. F.; Dohner, H.; Holzmann, K.; Dohner, K.; Bullinger, L., Integrative nucleophosmin mutation-associated microRNA and gene expression pattern analysis identifies novel microRNA - target gene interactions in acute myeloid leukemia. *Haematologica* **2011**, *96* (12), 1783-91.
241. Lin, L. T.; Chang, C. Y.; Chang, C. H.; Wang, H. E.; Chiou, S. H.; Liu, R. S.; Lee, T. W.; Lee, Y. J., Involvement of let-7 microRNA for the therapeutic effects of Rhenium-188-embedded liposomal nanoparticles on orthotopic human head and neck cancer model. *Oncotarget* **2016**, *7* (40), 65782-65796.
242. Vaz, C.; Ahmad, H. M.; Sharma, P.; Gupta, R.; Kumar, L.; Kulshreshtha, R.; Bhattacharya, A., Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. *BMC Genomics* **2010**, *11*, 288.
243. Spolverini, A.; Fuchs, G.; Bublik, D. R.; Oren, M., let-7b and let-7c microRNAs promote histone H2B ubiquitylation and inhibit cell migration by targeting multiple components of the H2B deubiquitylation machinery. *Oncogene* **2017**, *36* (42), 5819-5828.
244. Cheong, W. A. MicroRNA let-7a regulates integrin beta-3, vav3, and dicer to modulate trophoblast activities and hence embryo implantation. University of Hong Kong, Hong Kong, 2013.
245. Melton, C.; Judson, R. L.; Blelloch, R., Opposing microRNA families regulate self-renewal in mouse embryonic stem cells (vol 463, pg 621, 2010). *Nature* **2010**, *464* (7285), 126-126.
246. Patterson, M.; Gaeta, X.; Loo, K.; Edwards, M.; Smale, S.; Cinkornpumin, J.; Xie, Y.; Listgarten, J.; Azghadi, S.; Douglass, S. M.; Pellegrini, M.; Lowry, W. E., let-7 miRNAs Can Act through Notch to Regulate Human Gliogenesis. *Stem Cell Rep* **2014**, *3* (5), 758-773.
247. Diecke, S.; Quiroga-Negreira, A.; Redmer, T.; Besser, D., FGF2 signaling in mouse embryonic fibroblasts is crucial for self-renewal of embryonic stem cells. *Cells Tissues Organs* **2008**, *188* (1-2), 52-61.
248. Bhattacharya, B.; Cai, J.; Luo, Y.; Miura, T.; Mejido, J.; Brimble, S. N.; Zeng, X.; Schulz, T. C.; Rao, M. S.; Puri, R. K., Comparison of the gene expression profile of undifferentiated human embryonic stem cell lines and differentiating embryoid bodies. *BMC Dev Biol* **2005**, *5*, 22.
249. Lenz, M.; Goetzke, R.; Schenk, A.; Schubert, C.; Veeck, J.; Hemeda, H.; Koschmieder, S.; Zenke, M.; Schuppert, A.; Wagner, W., Epigenetic Biomarker to Support Classification into Pluripotent and Non-Pluripotent Cells. *Sci Rep-Uk* **2015**, *5*.
250. Liao, T. T.; Hsu, W. H.; Ho, C. H.; Hwang, W. L.; Lan, H. Y.; Lo, T.; Chang, C. C.; Tai, S. K.; Yang, M. H., let-7 Modulates Chromatin Configuration and Target Gene Repression through Regulation of the ARID3B Complex. *Cell Reports* **2016**, *14* (3), 520-533.
251. Zeng, X. M.; Miura, T.; Luo, Y. Q.; Bhattacharya, B.; Condie, B.; Chen, J.; Ginis, I.; Lyons, I.; Mejido, J.; Puri, R. K.; Rao, M. S.; Freed, W. J., Properties of pluripotent human embryonic stem cells BG01 and BG02. *Stem Cells* **2004**, *22* (3), 292-312.
252. Zhu, W.; Zhao, M.; Mattapally, S.; Chen, S.; Zhang, J., CCND2 Overexpression Enhances the Regenerative Potency of Human Induced Pluripotent Stem Cell-Derived Cardiomyocytes: Remuscularization of Injured Ventricle. *Circ Res* **2018**, *122* (1), 88-96.
253. Li, L. J.; Chen, Z. B.; Zhang, L. C.; Liu, G. Y.; Hua, J. L.; Jia, L. H.; Liao, M. Z., Genome-wide targets identification of "core" pluripotency transcription factors with integrated features in human embryonic stem cells. *Mol Biosyst* **2016**, *12* (4), 1324-1332.
254. Tetzlaff, M. T.; Bai, C.; Finegold, M.; Wilson, J.; Harper, J. W.; Mahon, K. A.; Elledge, S. J., Cyclin F disruption compromises placental development and affects normal cell cycle execution. *Mol Cell Biol* **2004**, *24* (6), 2487-98.
255. Zhang, X.; Neganova, I.; Przyborski, S.; Yang, C. B.; Cooke, M.; Atkinson, S. P.; Anyfantis, G.; Fenyk, S.; Keith, W. N.; Hoare, S. F.; Hughes, O.; Strachan, T.; Stojkovic, M.; Hinds, P. W.; Armstrong, L.; Lako, M., A role for NANOG in G1 to S transition in human embryonic stem cells through direct binding of CDK6 and CDC25A. *J Cell Biol* **2009**, *184* (1), 67-82.
256. Gaspar-Maia, A.; Alajem, A.; Polesso, F.; Sridharan, R.; Mason, M. J.; Heidersbach, A.; Ramalho-Santos, J.; McManus, M. T.; Plath, K.; Meshorer, E.; Ramalho-Santos, M., Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **2009**, *460* (7257), 863-8.
257. Wan, L.; Hu, X. J.; Yan, S. X.; Chen, F.; Cai, B.; Zhang, X. M.; Wang, T.; Yu, X. B.; Xiang, A. P.; Li, W. Q., Generation and neuronal differentiation of induced pluripotent stem cells in Cdy1^{-/-} mice. *Neuroreport* **2013**, *24* (3), 114-119.
258. Boland, M. J.; Nazor, K. L.; Loring, J. F., Epigenetic Regulation of Pluripotency and Differentiation. *Circulation Research* **2014**, *115* (2), 311-324.

259. Kumar, S.; Curran, J. E.; Glahn, D. C.; Blangero, J., Utility of Lymphoblastoid Cell Lines for Induced Pluripotent Stem Cell Generation. *Stem Cells Int* **2016**, 2016, 2349261.
260. Becker, K. A.; Stein, J. L.; Lian, J. B.; van Wijnen, A. J.; Stein, G. S., Establishment of histone gene regulation and cell cycle checkpoint control in human embryonic stem cells. *J Cell Physiol* **2007**, 210 (2), 517-26.
261. Richards, M.; Tan, S. P.; Tan, J. H.; Chan, W. K.; Bongso, A., The transcriptome profile of human embryonic stem cells as defined by SAGE. *Stem Cells* **2004**, 22 (1), 51-64.
262. Conway, A. E.; Van Nostrand, E. L.; Pratt, G. A.; Aigner, S.; Wilbert, M. L.; Sundararaman, B.; Freese, P.; Lambert, N. J.; Sathe, S.; Liang, T. Y.; Essex, A.; Landais, S.; Burge, C. B.; Jones, D. L.; Yeo, G. W., Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival. *Cell Reports* **2016**, 15 (3), 666-679.
263. Bell, J. L.; Wachter, K.; Muhleck, B.; Pazaitis, N.; Kohn, M.; Lederer, M.; Huttelmaier, S., Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell Mol Life Sci* **2013**, 70 (15), 2657-75.
264. Vogt, E. J.; Meglicki, M.; Hartung, K. I.; Borsuk, E.; Behr, R., Importance of the pluripotency factor LIN28 in the mammalian nucleolus during early embryonic development. *Development* **2012**, 139 (24), 4514-4523.
265. Zhang, J.; Ratanasirinawoot, S.; Chandrasekaran, S.; Wu, Z.; Ficarro, S. B.; Yu, C.; Ross, C. A.; Cacchiarelli, D.; Xia, Q.; Seligson, M.; Shinoda, G.; Xie, W.; Cahan, P.; Wang, L.; Ng, S. C.; Tintara, S.; Trapnell, C.; Onder, T.; Loh, Y. H.; Mikkelsen, T.; Sliz, P.; Teitell, M. A.; Asara, J. M.; Marto, J. A.; Li, H.; Collins, J. J.; Daley, G. Q., LIN28 Regulates Stem Cell Metabolism and Conversion to Primed Pluripotency. *Cell Stem Cell* **2016**, 19 (1), 66-80.
266. Armstrong, L.; Hughes, O.; Yung, S.; Hyslop, L.; Stewart, R.; Wappler, I.; Peters, H.; Walter, T.; Stojkovic, P.; Evans, J.; Stojkovic, M.; Lako, M., The role of PI3K/AKT, MAPK/ERK and NFkappabeta signalling in the maintenance of human embryonic stem cell pluripotency and viability highlighted by transcriptional profiling and functional analysis. *Hum Mol Genet* **2006**, 15 (11), 1894-913.
267. Gonzalez, F.; Huangfu, D., Mechanisms underlying the formation of induced pluripotent stem cells. *Wiley Interdiscip Rev Dev Biol* **2016**, 5 (1), 39-65.
268. Li, L.; Walsh, R. M.; Wagh, V.; James, M. F.; Beauchamp, R. L.; Chang, Y. S.; Gusella, J. F.; Hochedlinger, K.; Ramesh, V., Mediator Subunit Med28 Is Essential for Mouse Peri-Implantation Development and Pluripotency. *Plos One* **2015**, 10 (10).
269. Varlakhanova, N. V.; Cotterman, R. F.; deVries, W. N.; Morgan, J.; Donahue, L. R.; Murray, S.; Knowles, B. B.; Knoepfler, P. S., myc maintains embryonic stem cell pluripotency and self-renewal. *Differentiation* **2010**, 80 (1), 9-19.
270. Yan, Y.; Yin, P. P.; Gong, H.; Xue, Y. Y.; Zhang, G. P.; Fang, B.; Chen, Z. D.; Li, Y.; Yang, C. J.; Huang, Z. Y.; Yang, X. D.; Ge, J. B.; Zou, Y. Z., Nucleosome Assembly Protein 1-Like 1 (Nap1l1) Regulates the Proliferation of Murine Induced Pluripotent Stem Cells. *Cell Physiol Biochem* **2016**, 38 (1), 340-350.
271. Zhu, Y.; Liu, H.; Xu, L.; An, H.; Liu, W.; Liu, Y.; Lin, Z.; Xu, J., p21-activated kinase 1 determines stem-like phenotype and sunitinib resistance via NF-kappaB/IL-6 activation in renal cell carcinoma. *Cell Death Dis* **2015**, 6, e1637.
272. Zheng, H. W.; Ying, H. Q.; Wiedemeyer, R.; Yan, H. Y.; Quayle, S. N.; Ivanova, E. V.; Paik, J. H.; Zhang, H. L.; Xiao, Y. H.; Perry, S. R.; Hu, J.; Vinjamoori, A.; Gan, B. Y.; Sahin, E.; Chheda, M. G.; Brennan, C.; Wang, Y. A.; Hahn, W. C.; Chin, L.; DePinho, R. A., PLAGL2 Regulates Wnt Signaling to Impede Differentiation in Neural Stem Cells and Gliomas. *Cancer Cell* **2010**, 17 (5), 497-509.
273. Nishino, K.; Toyoda, M.; Yamazaki-Inoue, M.; Makino, H.; Fukawatase, Y.; Chikazawa, E.; Takahashi, Y.; Miyagawa, Y.; Okita, H.; Kiyokawa, N.; Akutsu, H.; Umezawa, A., Defining hypo-methylated regions of stem cell-specific promoters in human iPS cells derived from extra-embryonic amnions and lung fibroblasts. *PLoS One* **2010**, 5 (9), e13017.
274. Pripuzova, N. S.; Getie-Kehtie, M.; Grunseich, C.; Sweeney, C.; Malech, H.; Alterman, M. A., Development of a protein marker panel for characterization of human induced pluripotent stem cells (hiPSCs) using global quantitative proteome analysis. *Stem Cell Res* **2015**, 14 (3), 323-338.
275. Awe, J. P.; Crespo, A. V.; Li, Y.; Kiledjian, M.; Byrne, J. A., BAY11 enhances OCT4 synthetic mRNA expression in adult human skin cells. *Stem Cell Res Ther* **2013**, 4 (1), 15.
276. Xiao, S.; Lu, J.; Sridhar, B.; Cao, X.; Yu, P.; Zhao, T.; Chen, C. C.; McDee, D.; Sloofman, L.; Wang, Y.; Rivas-Astroza, M.; Telugu, B.; Lefebvre, D.; Zhang, K.; Liang, H.; Zhao, J. C.; Tanaka, T. S.; Stormo, G.; Zhong, S., SMARCD1 Contributes to the Regulation of Naive Pluripotency by Interacting with Histone Citrullination. *Cell Rep* **2017**, 18 (13), 3117-3128.

277. Huang, X.; Tian, C. H.; Liu, M.; Wang, Y. X.; Tolmachev, A. V.; Sharma, S.; Yu, F.; Fu, K.; Zheng, J. L.; Ding, S. J., Quantitative Proteomic Analysis of Mouse Embryonic Fibroblasts and Induced Pluripotent Stem Cells Using O-16/O-18 Labeling. *J Proteome Res* **2012**, *11* (4), 2091-2102.
278. Baumann, K., Stem cells: TFIID promotes pluripotency. *Nat Rev Mol Cell Biol* **2013**, *14* (5), 264.
279. Tropel, P.; Jung, L.; Andre, C.; Ndandougou, A.; Viville, S., CpG Island Methylation Correlates with the Use of Alternative Promoters for USP44 Gene Expression in Human Pluripotent Stem Cells and Testes. *Stem Cells Dev* **2017**, *26* (15), 1100-1110.
280. Fuchs, G.; Shema, E.; Vesterman, R.; Kotler, E.; Wolchinsky, Z.; Wilder, S.; Golomb, L.; Pribluda, A.; Zhang, F.; Haj-Yahya, M.; Feldmesser, E.; Brik, A.; Yu, X.; Hanna, J.; Aberdam, D.; Domany, E.; Oren, M., RNF20 and USP44 regulate stem cell differentiation by modulating H2B monoubiquitylation. *Mol Cell* **2012**, *46* (5), 662-73.
281. Liu, R. L.; Li, J.; Lai, Y. H.; Liao, Y.; Liu, R. M.; Qiu, W. S., Hsa-miR-1 suppresses breast cancer development by down-regulating K-ras and long non-coding RNA MALAT1. *Int J Biol Macromol* **2015**, *81*, 491-497.
282. Yan, L. X.; Huang, X. F.; Shao, Q.; Huang, M. Y.; Deng, L.; Wu, Q. L.; Zeng, Y. X.; Shao, J. Y., MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *Rna-a Publication of the Rna Society* **2008**, *14* (11), 2348-2360.
283. Marcucci, G.; Mrozek, K.; Radmacher, M. D.; Garzon, R.; Bloomfield, C. D., The prognostic and functional role of microRNAs in acute myeloid leukemia. *Blood* **2011**, *117* (4), 1121-9.
284. Zhu, C.; Wang, Y.; Kuai, W.; Sun, X.; Chen, H.; Hong, Z., Prognostic value of miR-29a expression in pediatric acute myeloid leukemia. *Clin Biochem* **2013**, *46* (1-2), 49-53.
285. Wu, Z.; Huang, X.; Huang, X.; Zou, Q.; Guo, Y., The inhibitory role of Mir-29 in growth of breast cancer cells. *J Exp Clin Cancer Res* **2013**, *32*, 98.
286. Zhang, X. W.; Zhao, X. H.; Fiskus, W.; Lin, J. H.; Lwin, T.; Rao, R.; Zhang, Y. Z.; Chan, J. C.; Fu, K.; Marquez, V. E.; Chen-Kiang, S.; Moscinski, L. C.; Seto, E.; Dalton, W. S.; Wright, K. L.; Sotomayor, E.; Bhalla, K.; Tao, J. G., Coordinated Silencing of MYC-Mediated miR-29 by HDAC3 and EZH2 as a Therapeutic Target of Histone Modification in Aggressive B-Cell Lymphomas. *Cancer Cell* **2012**, *22* (4), 506-523.
287. Xi, Z.; Wang, P.; Xue, Y.; Shang, C.; Liu, X.; Ma, J.; Li, Z.; Li, Z.; Bao, M.; Liu, Y., Overexpression of miR-29a reduces the oncogenic properties of glioblastoma stem cells by downregulating Quaking gene isoform 6. *Oncotarget* **2017**, *8* (15), 24949-24963.
288. Garzon, R.; Heaphy, C. E.; Havelange, V.; Fabbri, M.; Volinia, S.; Tsao, T.; Zanesi, N.; Kornblau, S. M.; Marcucci, G.; Calin, G. A.; Andreeff, M.; Croce, C. M., MicroRNA 29b functions in acute myeloid leukemia. *Blood* **2009**, *114* (26), 5331-41.
289. Drago-Ferrante, R.; Pentimalli, F.; Carlisi, D.; De Blasio, A.; Saliba, C.; Baldacchino, S.; Degaetano, J.; Debono, J.; Caruana-Dingli, G.; Grech, G.; Scerri, C.; Tesoriere, G.; Giordano, A.; Vento, R.; Di Fiore, R., Suppressive role exerted by microRNA-29b-1-5p in triple negative breast cancer through SPIN1 regulation. *Oncotarget* **2017**, *8* (17), 28939-28958.
290. Catania, A.; Maira, F.; Skarmoutsou, E.; D'Amico, F.; Abounader, R.; Mazzarino, M. C., Insight into the role of microRNAs in brain tumors (Review). *International Journal of Oncology* **2012**, *40* (3), 605-624.
291. Li, W.; Yi, J.; Zheng, X. J.; Liu, S. W.; Fu, W. Q.; Ren, L. W.; Li, L.; Hoon, D. S. B.; Wang, J. H.; Du, G. H., miR-29c plays a suppressive role in breast cancer by targeting the TIMP3/STAT1/FOXO1 pathway. *Clin Epigenetics* **2018**, *10*.
292. Li, L. S.; Yuan, L. J.; Luo, J. M.; Gao, J.; Guo, J. L.; Xie, X. M., MiR-34a inhibits proliferation and migration of breast cancer through down-regulation of Bcl-2 and SIRT1. *Clin Exp Med* **2013**, *13* (2), 109-117.
293. Liu, Y. P.; Hu, H.; Xu, F.; Wen, J. J., [Relation of MiR-34a Expression in Diffuse Large B Cell Lymphoma with Clinical Prognosis]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi* **2017**, *25* (2), 455-459.
294. Romero, P. V.; Cialfi, S.; Palermo, R.; De Blasio, C.; Checquolo, S.; Bellavia, D.; Chiaretti, S.; Foa, R.; Amadori, A.; Gulino, A.; Zardo, G.; Talora, C.; Screpanti, I., The deregulated expression of miR-125b in acute myeloid leukemia is dependent on the transcription factor C/EBP alpha. *Leukemia* **2015**, *29* (12), 2442-2445.
295. Caramuta, S.; Lee, L.; Ozata, D. M.; Akcakaya, P.; Georgii-Hemming, P.; Xie, H.; Amini, R. M.; Lawrie, C. H.; Enblad, G.; Larsson, C.; Berglund, M.; Lui, W. O., Role of microRNAs and microRNA machinery in the pathogenesis of diffuse large B-cell lymphoma. *Blood Cancer J* **2013**, *3*, e152.
296. Zheng, M. Z.; Sun, X.; Li, Y. Q.; Zuo, W. S., MicroRNA-145 inhibits growth and migration of breast cancer cells through targeting oncoprotein ROCK1. *Tumor Biol* **2016**, *37* (6), 8189-8196.
297. Bradshaw, G.; Sutherland, H. G.; Haupt, L. M.; Griffiths, L. R., Dysregulated MicroRNA Expression Profiles and Potential Cellular, Circulating and Polymorphic Biomarkers in Non-Hodgkin Lymphoma. *Genes-Basel* **2016**, *7* (12).

298. Mattiske, S.; Suetani, R. J.; Neilsen, P. M.; Callen, D. F., The Oncogenic Role of miR-155 in Breast Cancer. *Cancer Epidem Biomar* **2012**, *21* (8), 1236-1243.
299. Li, B.; Lu, Y.; Wang, H.; Han, X.; Mao, J.; Li, J.; Yu, L.; Wang, B.; Fan, S.; Yu, X.; Song, B., miR-221/222 enhance the tumorigenicity of human breast cancer stem cells via modulation of PTEN/Akt pathway. *Biomed Pharmacother* **2016**, *79*, 93-101.
300. Liu, X.; Lv, X. P.; Yang, Q. K.; Jin, H. F.; Zhou, W. P.; Fan, Q. X., MicroRNA-29a Functions as a Tumor Suppressor and Increases Cisplatin Sensitivity by Targeting NRAS in Lung Cancer. *Technol Cancer Res T* **2018**, *17*.
301. Chen, H. X.; Xu, X. X.; Zhang, Z.; Tan, B. Z.; Zhou, X. D., MicroRNA-29b Inhibits Angiogenesis by Targeting VEGFA through the MAPK/ERK and PI3K/Akt Signaling Pathways in Endometrial Carcinoma. *Cell Physiol Biochem* **2017**, *41* (3), 933-946.
302. Peng, Y.; Guo, J. J.; Liu, Y. M.; Wu, X. L., MicroRNA-34A inhibits the growth, invasion and metastasis of gastric cancer by targeting PDGFR and MET expression. *Bioscience Rep* **2014**, *34*, 247-256.
303. Xie, M.; Dart, D. A.; Guo, T.; Xing, X. F.; Cheng, X. J.; Du, H.; Jiang, W. G.; Wen, X. Z.; Ji, J. F., MicroRNA-1 acts as a tumor suppressor microRNA by inhibiting angiogenesis-related growth factors in human gastric cancer. *Gastric Cancer* **2018**, *21* (1), 41-54.
304. Han, C.; Zhou, Y. B.; An, Q.; Li, F.; Li, D. L.; Zhang, X. J.; Yu, Z. J.; Zheng, L. L.; Duan, Z. F.; Kan, Q. C., MicroRNA-1 (miR-1) inhibits gastric cancer cell proliferation and migration by targeting MET. *Tumor Biol* **2015**, *36* (9), 6715-6723.
305. Liu, S.; Gao, G. Z.; Yan, D. X.; Chen, X. J.; Yao, X. W.; Guo, S. Z.; Li, G. R.; Zhao, Y., Effects of miR-145-5p through NRAS on the cell proliferation, apoptosis, migration, and invasion in melanoma by inhibiting MAPK and PI3K/AKT pathways. *Cancer Med-Uls* **2017**, *6* (4), 819-833.

Acknowledgement (감사의 글)

학위 논문의 마지막 장을 쓰고 있는 지금 이 순간, 지난 6 년 간의 대학원 생활을 돌이켜보니 ‘참 정신없이 바쁘게 지내 왔구나’ 하는 생각과 함께 ‘그래도 더 열심히 살 수 있지 않았을까’ 하는 아쉬움이 남습니다. 그리고 앞으로 얻게 될 박사라는 타이틀의 무게에 대해서도 생각해보게 됩니다.

그동안 학위 과정을 무사히 마무리할 수 있도록 도움을 주신 분들이 많아, 마지막 장에서는 그분들에 대한 감사를 표하고자 합니다. 가장 먼저, 연구가 올바른 방향으로 진행될 수 있도록 아낌없는 조언과 가르침을 주신 남덕우 교수님께 감사드립니다. 부족한 점이 많은 저에게 그동안 많은 격려를 해주신 교수님 덕분에 힘을 내서 연구를 할 수 있었던 것 같습니다. 교수님의 연구에 대한 열정과 훌륭하신 인품을 본받아 앞으로 좋은 연구자로 거듭날 수 있도록 노력하겠습니다. 다음으로, 바쁘신 와중에도 귀한 시간을 내주시어 제 학위 논문을 심사해주시고 많은 조언을 해주신 김철민, 유연주, 권태준, 이세민 교수님께 감사드립니다. 특히, 유연주 교수님께 GWAS simulation modeling 을 도와주셔서 정말 감사드린다는 말씀 전하고 싶습니다. 그리고 miRNA-29 의 타겟 검증을 도와주신 박지영 교수님과 조우빈 학생에게도 감사를 전합니다.

늘 제가 어디 아픈 데는 없는지, 밥은 잘 먹고 다니는지 걱정해주고 격려해주었던 아빠, 엄마, 언니, 동생에게 감사를 포함합니다. 앞으로 건강 잘 챙길게요! 그리고 항상 저를 아껴주시고 응원해주신 친지 여러분께도 감사드립니다. 저희 실험실 멤버 분들에게도 고마움을 전하고 싶습니다. 우선 다년간의 유학 생활에서 얻은 풍부한 경험을 들려주시고, 제게 꿈을 심어주신 주옥언니에게 감사합니다. 그리고 언제나 웃는 얼굴로 반겨주는 ‘Genius’ Hai 박사님께 감사합니다. Java script/Shiny 마스터 김진환을 비롯한 후배 대학원생들아, 화이팅! 또한, 존재만으로도 힘이 되는 생명과 09 학번 절친들인 김영린, 신이슬, 이안중, 이유림, 임한솔, 최성열에게 고맙습니다. 생명과는 아니지만 다년간 룸메이트로 지내며 저에게 많은 정보와 좋은 영향을 주었던 친구 김선아에게도 감사를 포함합니다. 또 대학원 마지막 학기에 졸업 준비의 고통을 함께 나눈 현모오빠, 상호, 진영 언니, 정말 수고했어요!! 마지막으로, 항상 저와 함께 다니며 함께 고민도 나누고, 맛있는 음식도 나누고, 연구에 대해 토론도 하고, 힘들 때 제게 힘을 주었던 장준일에게 감사합니다. 앞으로도 지금처럼 함께하며 서로에게 좋은 영향을 주는 연구자로 성장해 나갔으면 좋겠습니다.

저희 지도 교수님께서 제게 종종 ‘연구자는 좋은 연구 논문으로 승부한다’는 말씀을 해주셨습니다. 이 말씀을 항상 명심하고, 인류 삶에 작은 부분이나마 공헌할 수 있는 연구자가 될 수 있도록 노력하겠습니다. 감사합니다.